



# **An Overview of High Performance Computing and Future Requirements**

---

**Jack Dongarra**

**University of Tennessee  
Oak Ridge National Laboratory**





Oak Ridge National Lab

x

x

University of Tennessee, Knoxville

# Overview

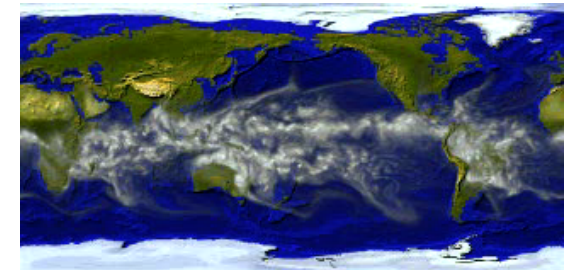
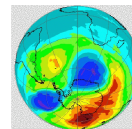
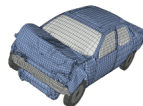
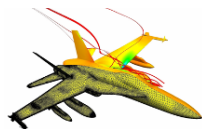
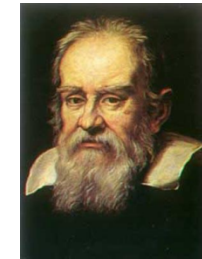
---

- **Computational Science**
- **High Performance Computing**
- **Projections for the Future**

# Simulation: The Third Pillar of Science



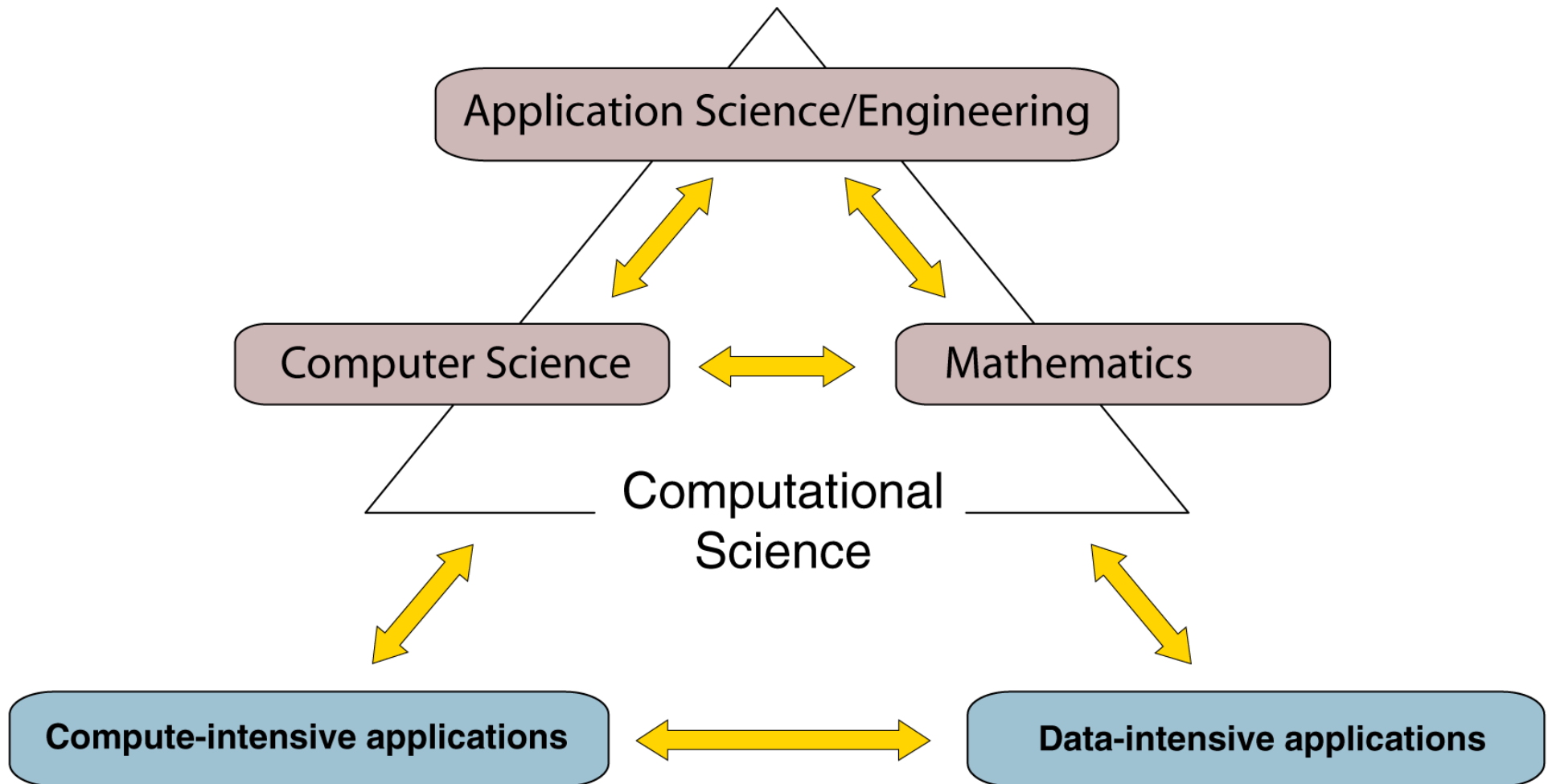
- Traditional scientific and engineering paradigm:
  - 1) Do theory or paper design.
  - 2) Perform experiments or build system.
- Limitations:
  - Too difficult -- build large wind tunnels.
  - Too expensive -- build a throw-away passenger jet.
  - Too slow -- wait for climate or galactic evolution.
  - Too dangerous -- weapons, drug design, climate experimentation.
- Computational science paradigm:
  - 3) Use high performance computer systems to simulate the phenomenon
    - Base on known physical laws and efficient numerical methods.



# Computational Science Fuses Three Distinct Elements:

---

5



# Look at the Fastest Computers

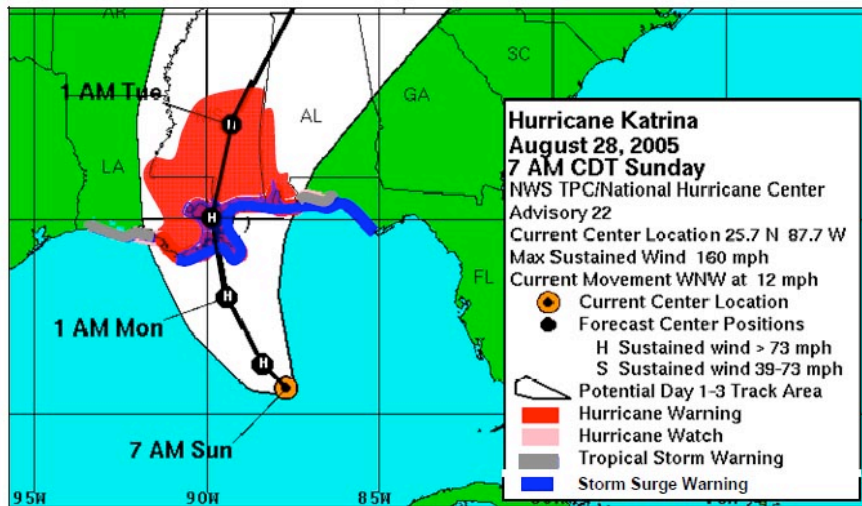
---

- **Supercomputing Matters**
  - Essential for scientific discovery
  - Critical for national security
  - Fundamental contributor to the economy and competitiveness through use in engineering and manufacturing
- Supercomputers are *the tool for solving the most challenging problems through simulations*



# Weather and Economic Loss

- 40% of the \$14T U.S. economy is impacted by weather and climate
- \$1M in economic loss to evacuate each 1 mile of coastline

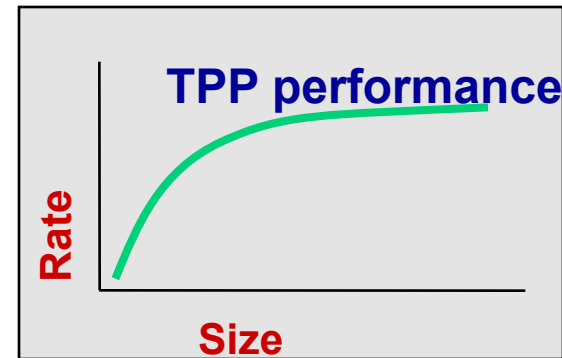


- We now over-warn by a factor of 3
- Average over-warning is 200 miles, or \$200M per event
- Improved forecasts
  - saving lives and resources

H. Meuer, H. Simon, E. Strohmaier, & JD

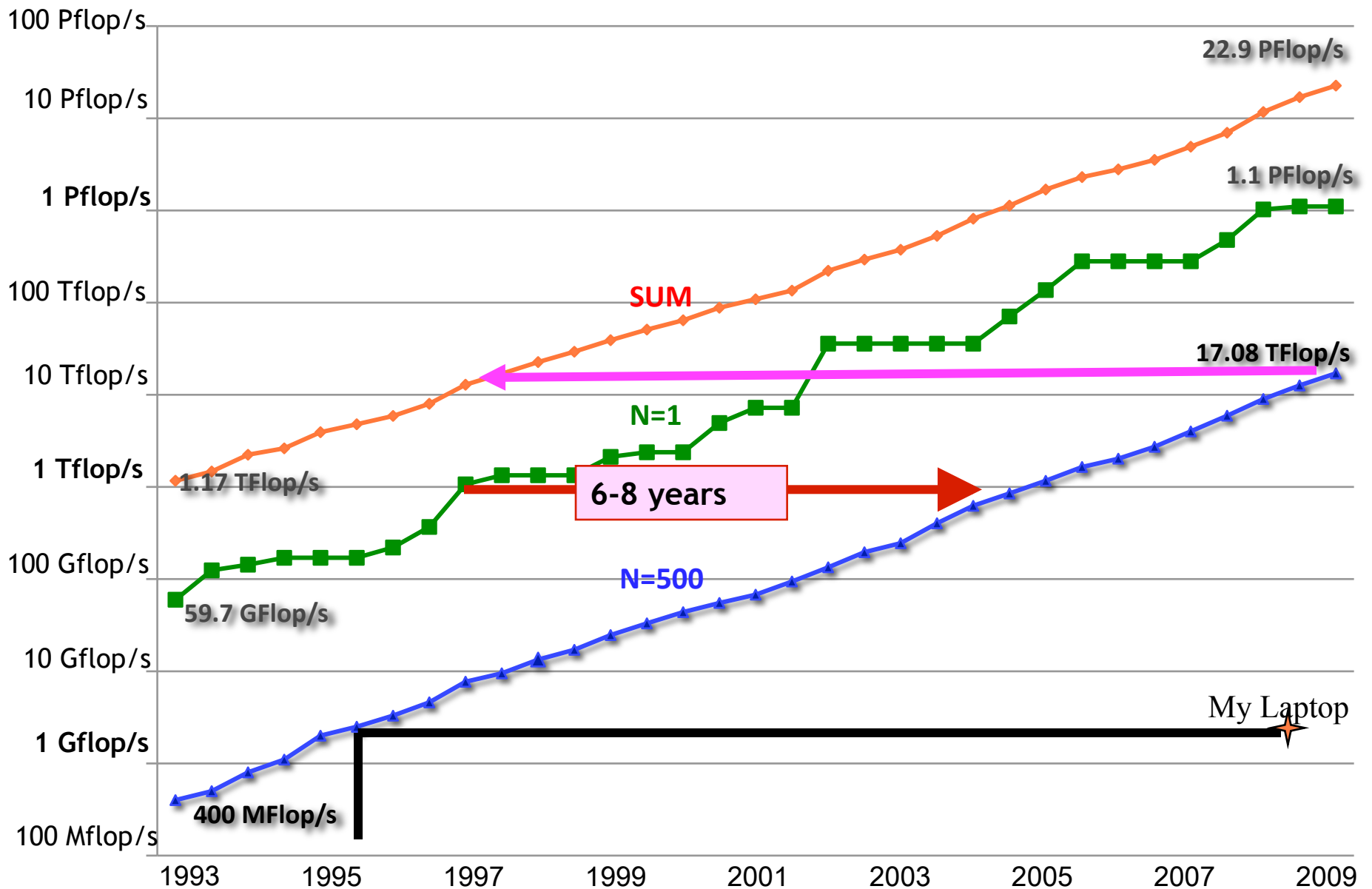
- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

$$Ax=b, \text{ dense problem}$$



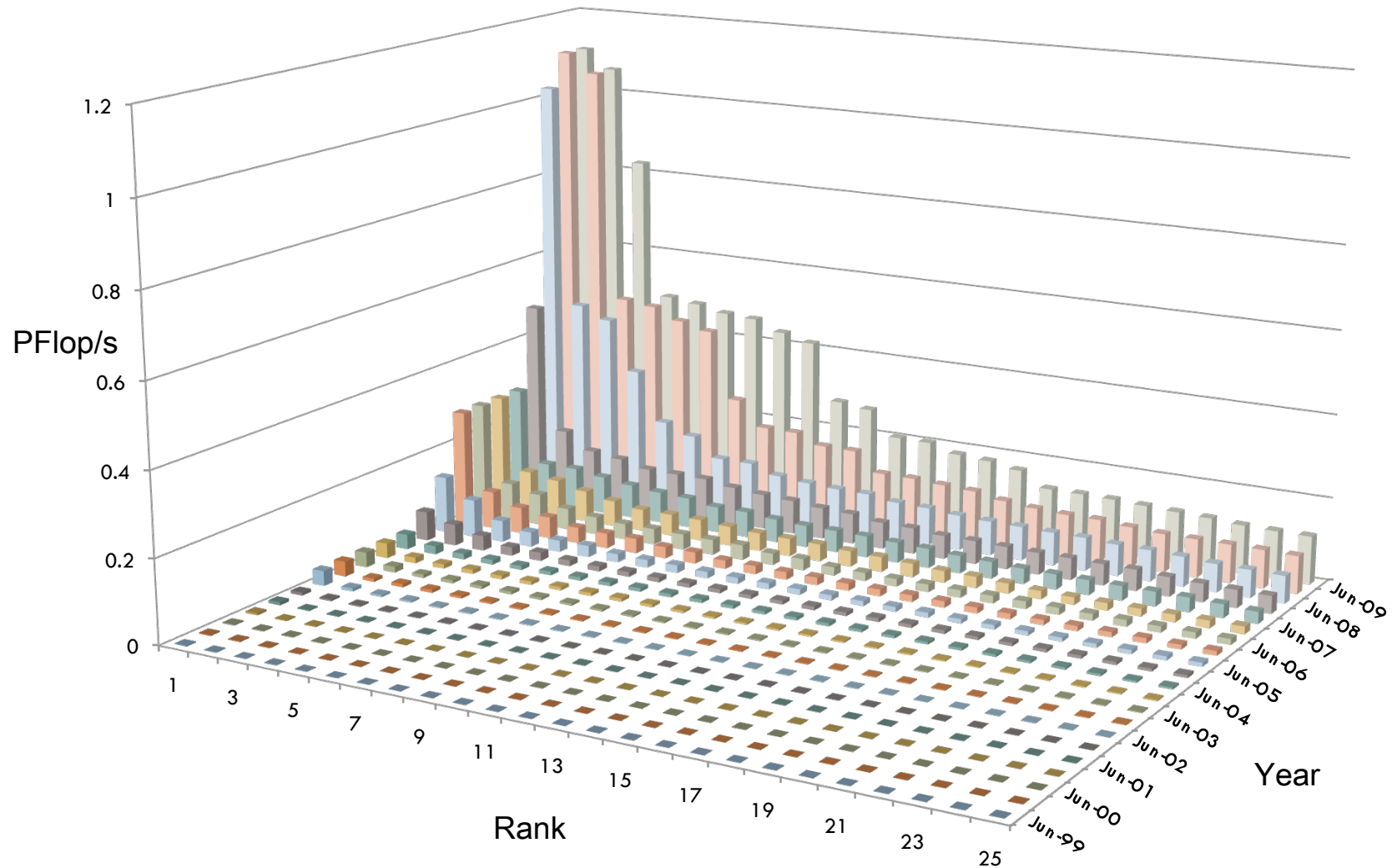
- Updated twice a year  
SC'xy in the States in November  
Meeting in Germany in June
- All data available from [www.top500.org](http://www.top500.org)

# Performance Development

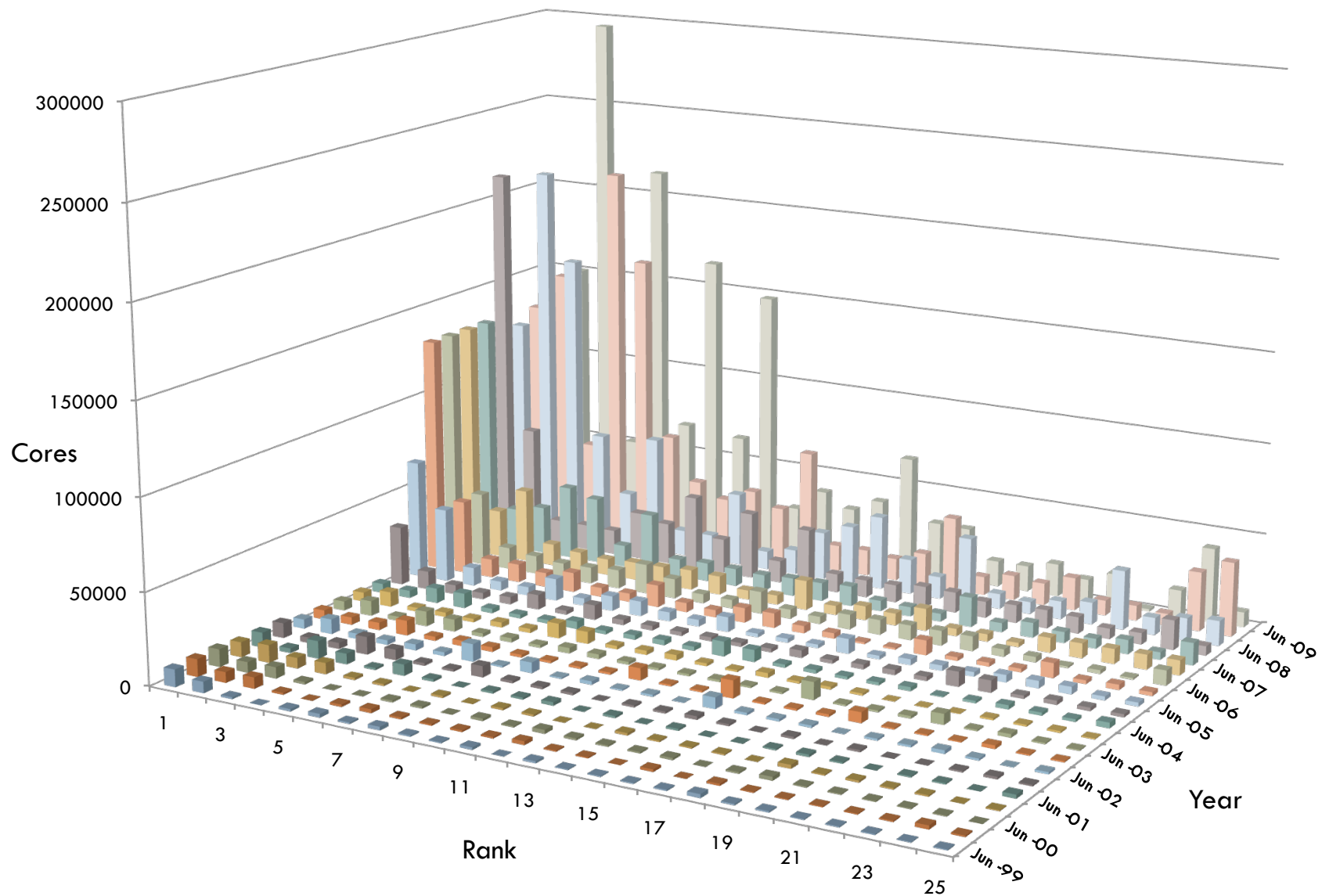




# Performance of Top25 Over 10 Years



# Cores in the Top25 Over Last 10 Years



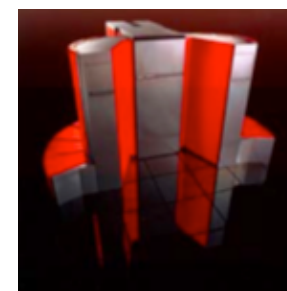


# Looking at the Gordon Bell Prize

(Recognize outstanding achievement in high-performance computing applications and encourage development of parallel processing )

- 1 GFlop/s; 1988; Cray Y-MP; 8 Processors

- ▣ Static finite element analysis



- 1 TFlop/s; 1998; Cray T3E; 1024 Processors

- ▣ Modeling of metallic magnet atoms, using a variation of the locally self-consistent multiple scattering method.



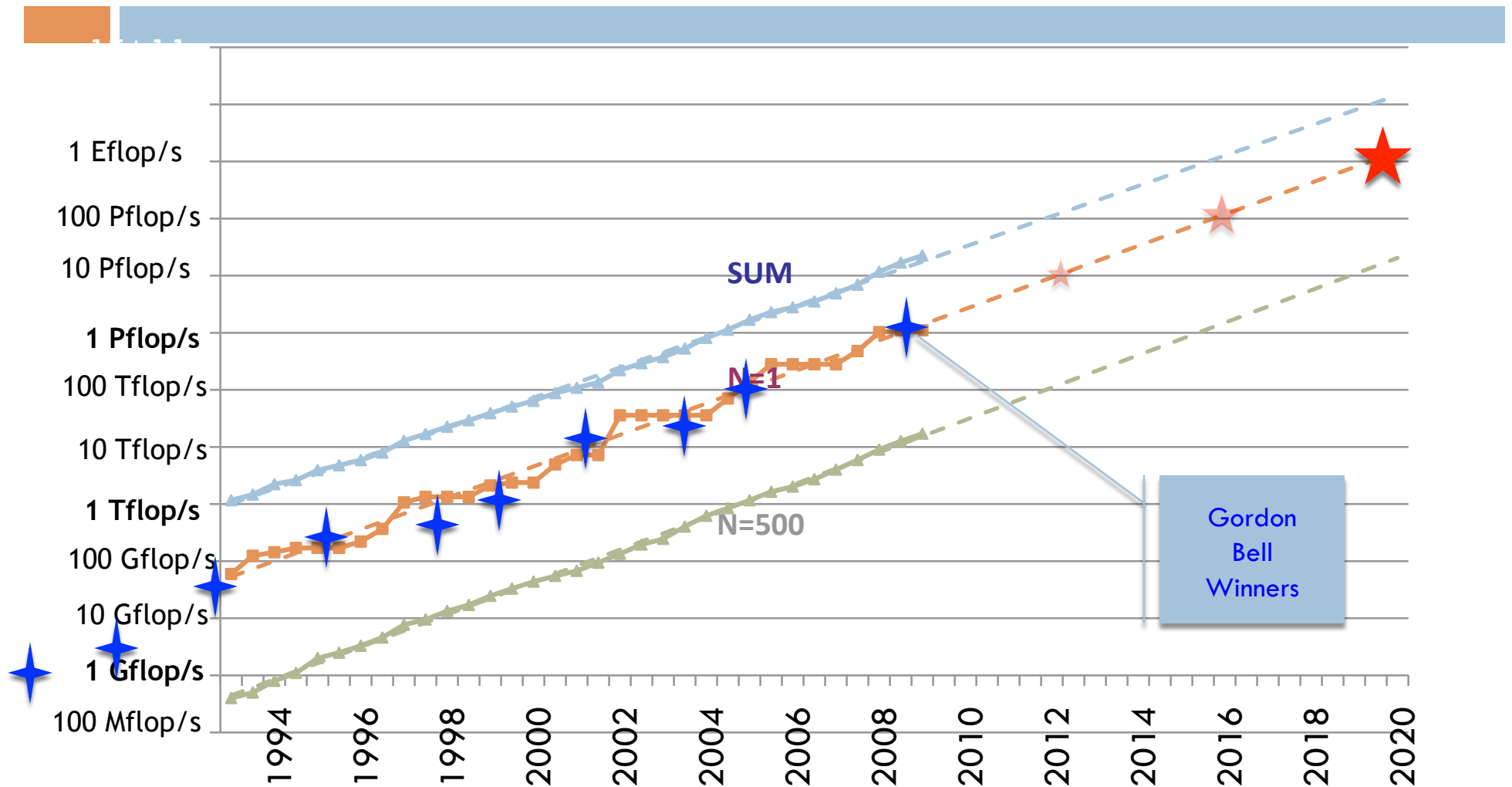
- 1 PFlop/s; 2008; Cray XT5;  $1.5 \times 10^5$  Processors

- ▣ Superconductive materials

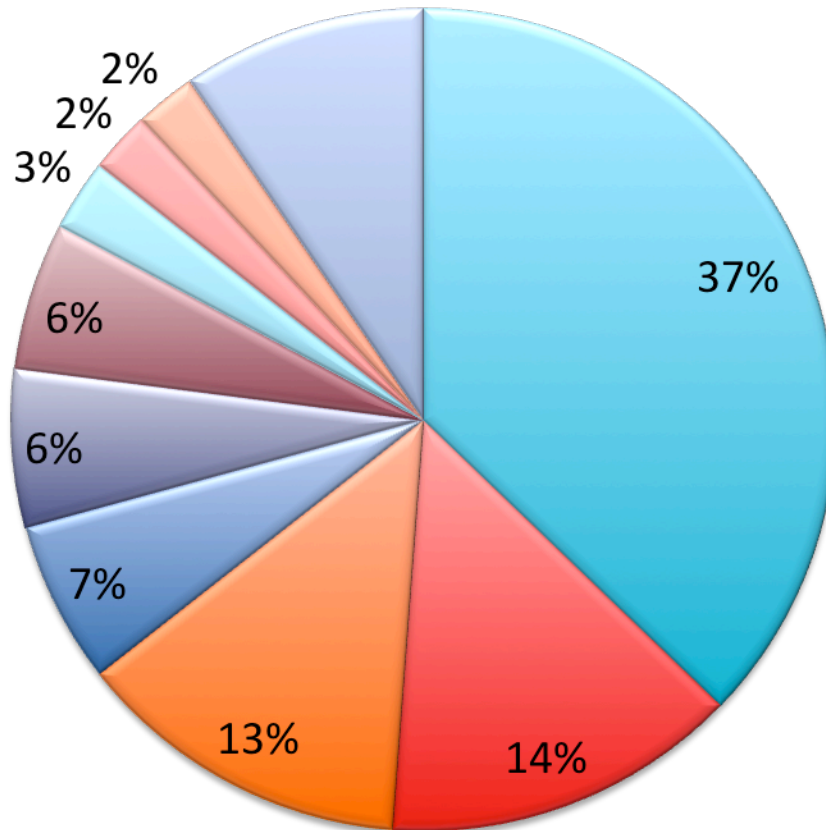


- 1 EFlop/s; ~2018; ?;  $1 \times 10^7$  Processors ( $10^9$  threads)

# Performance Development in Top500



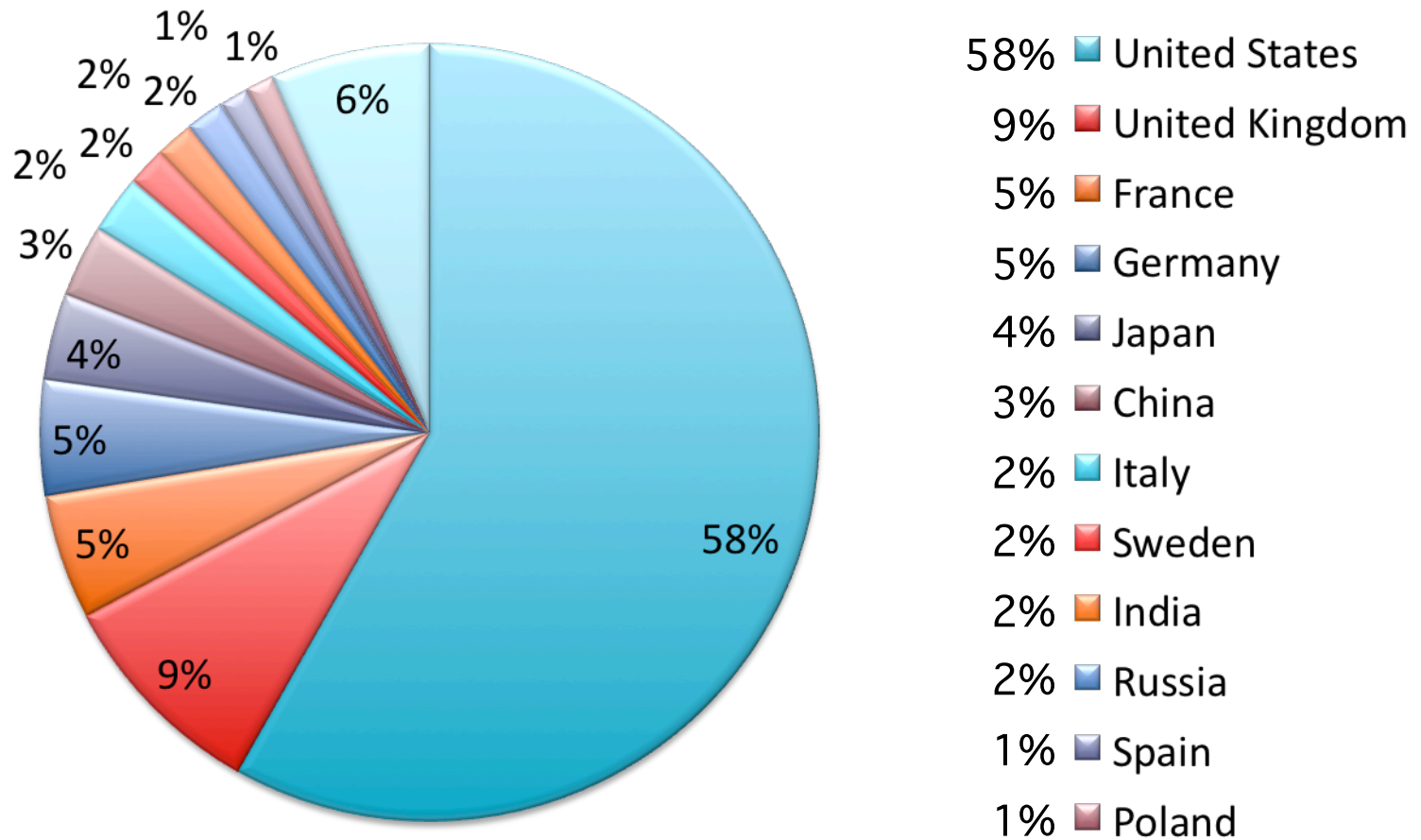
# Processors Used in Supercomputers



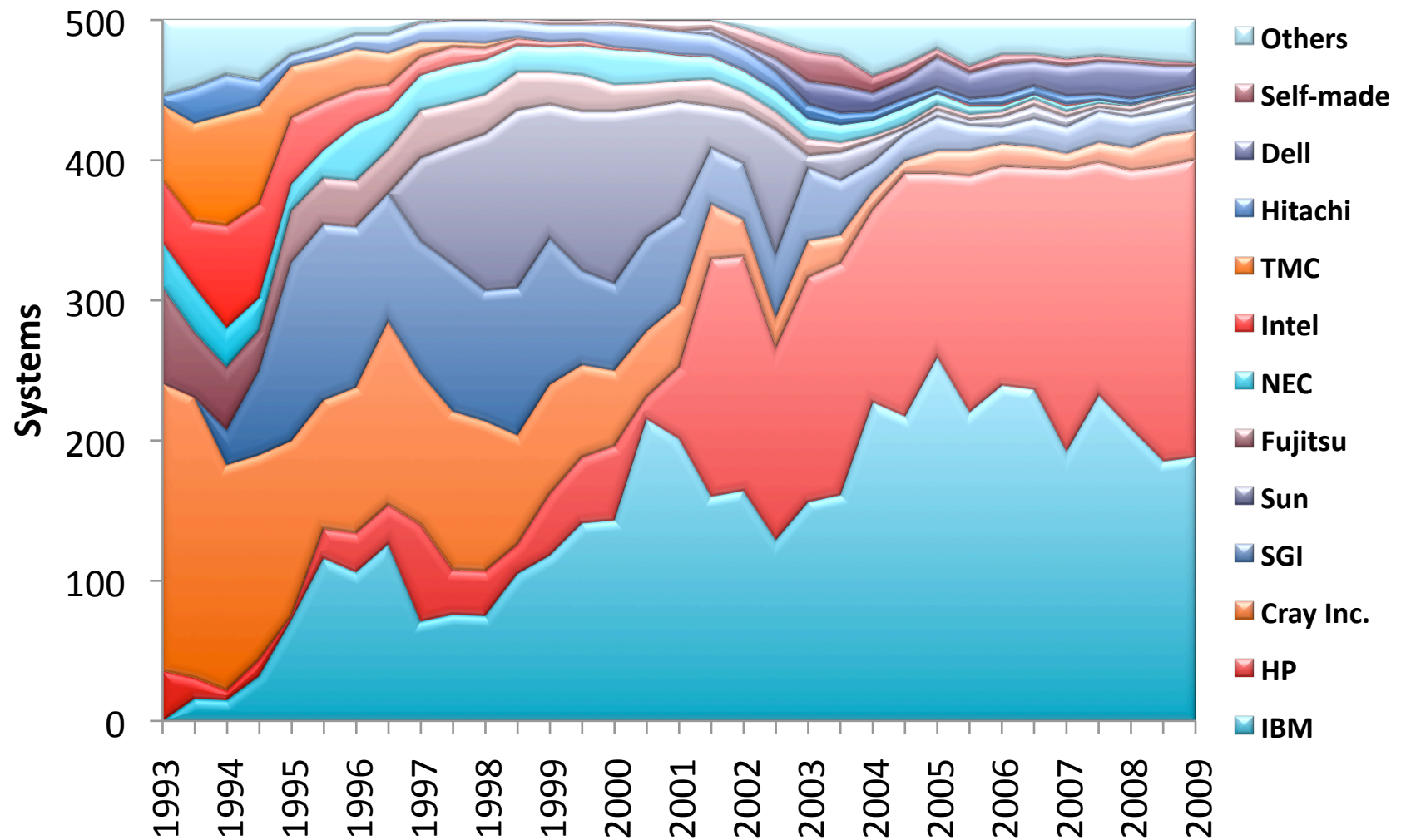
- Xeon E54xx (Harpertown)
- Xeon 51xx (Woodcrest)
- Xeon 53xx (Clovertown)
- Xeon L54xx (Harpertown)
- Opteron Quad Core
- Opteron Dual Core
- PowerPC 440
- PowerPC 450
- POWER6
- Others

Intel 71%  
AMD 13%  
IBM 7%

# Countries / System Share

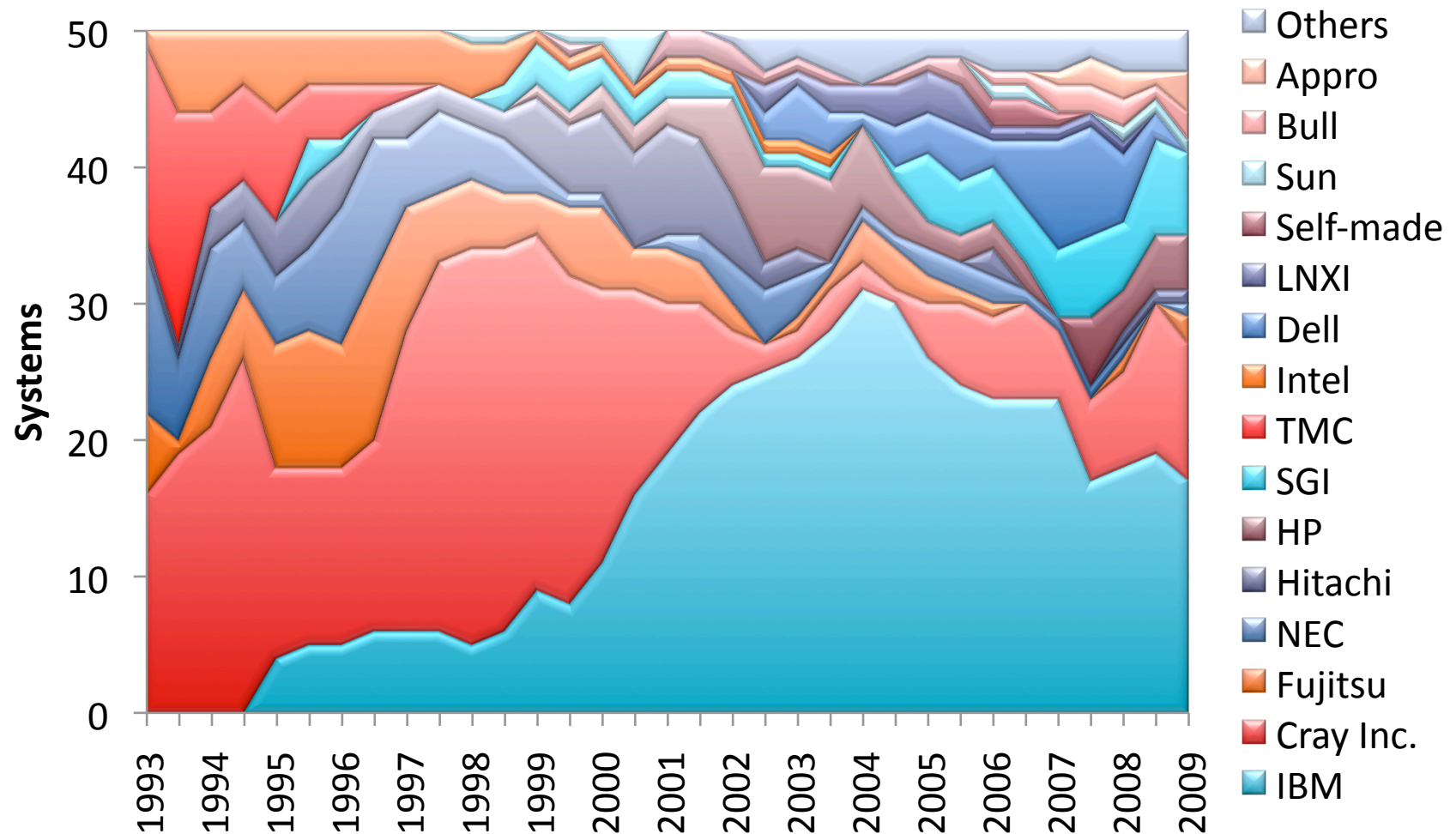


# Vendors



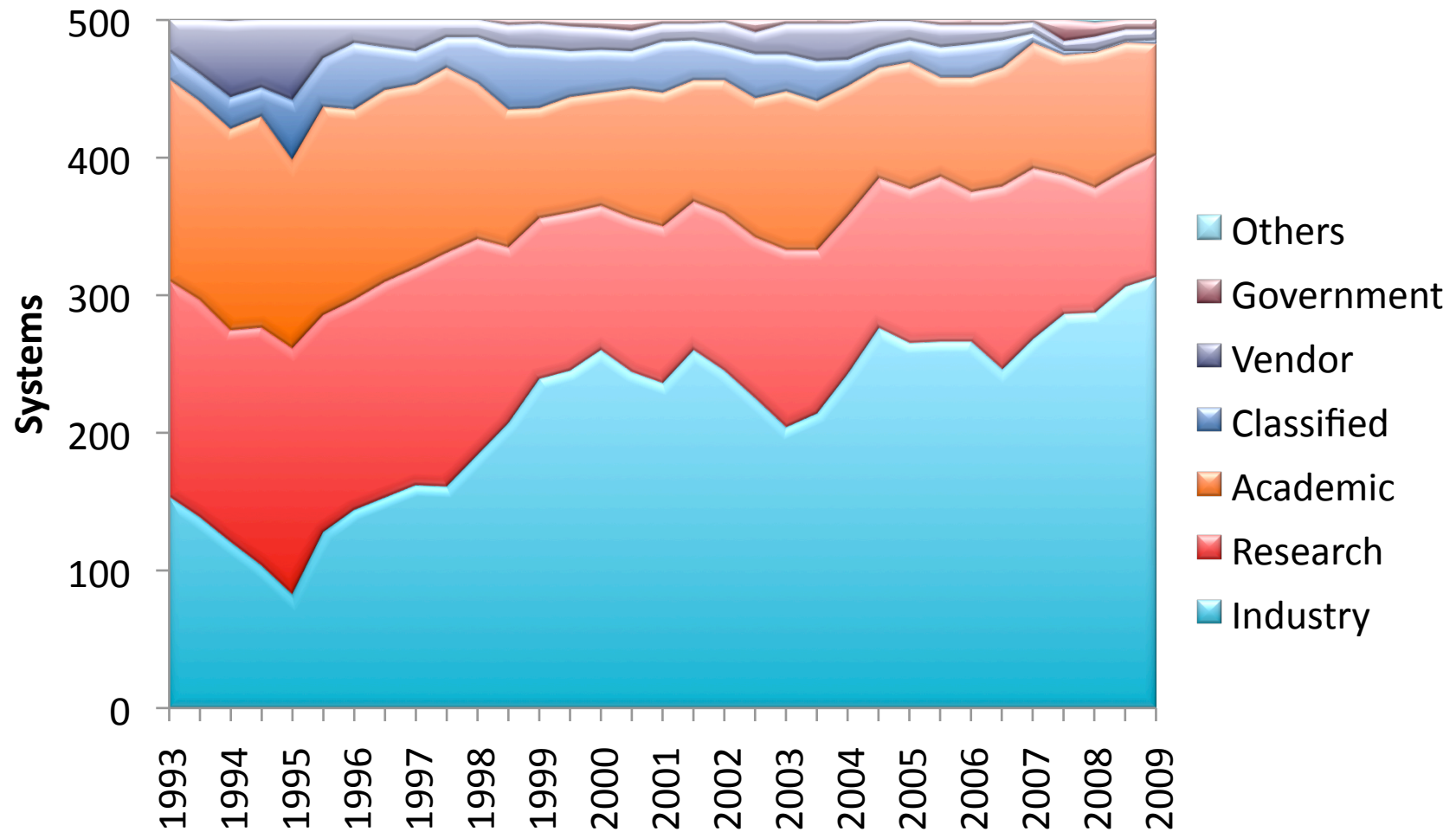


# Vendors (TOP50)

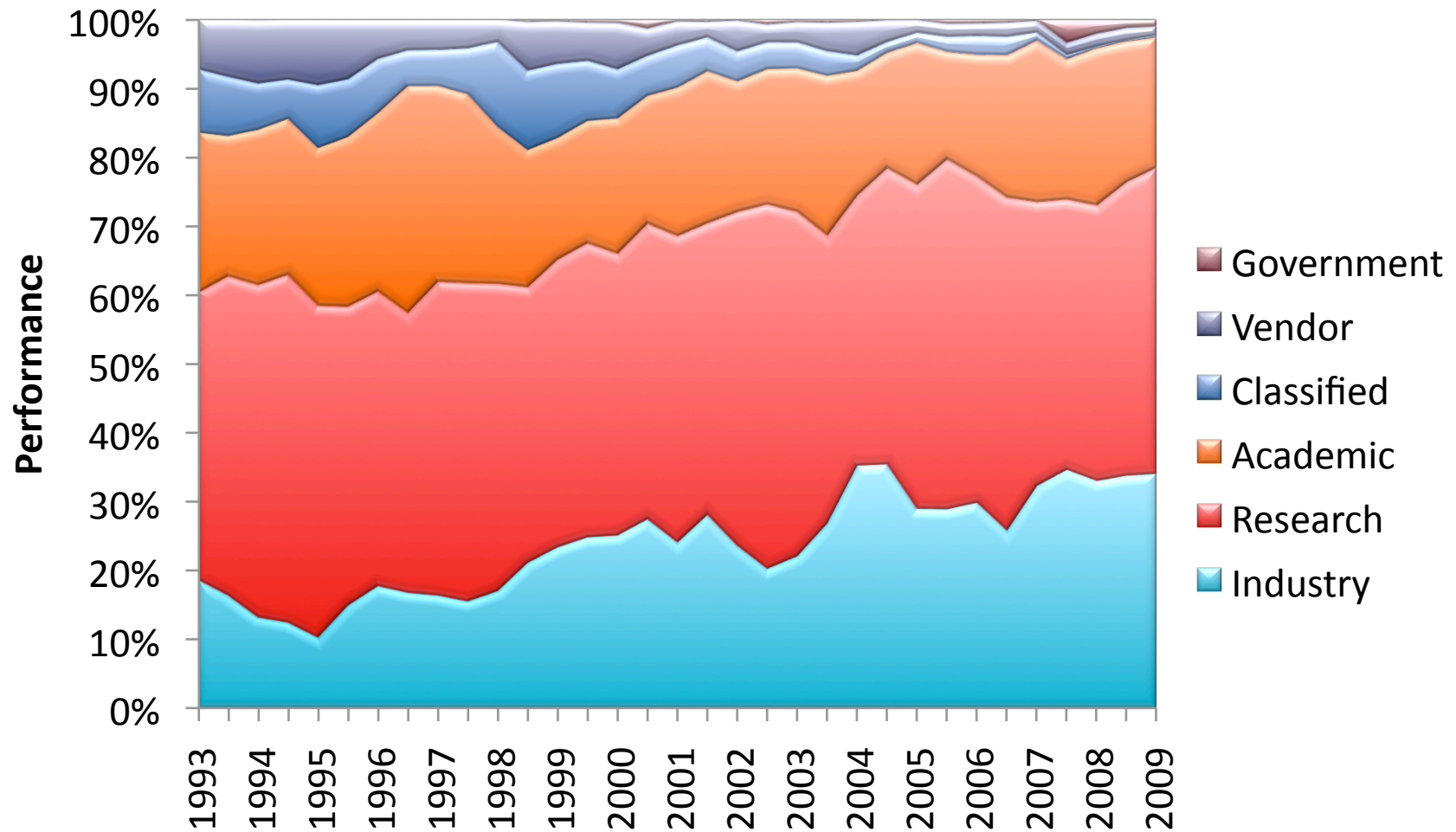




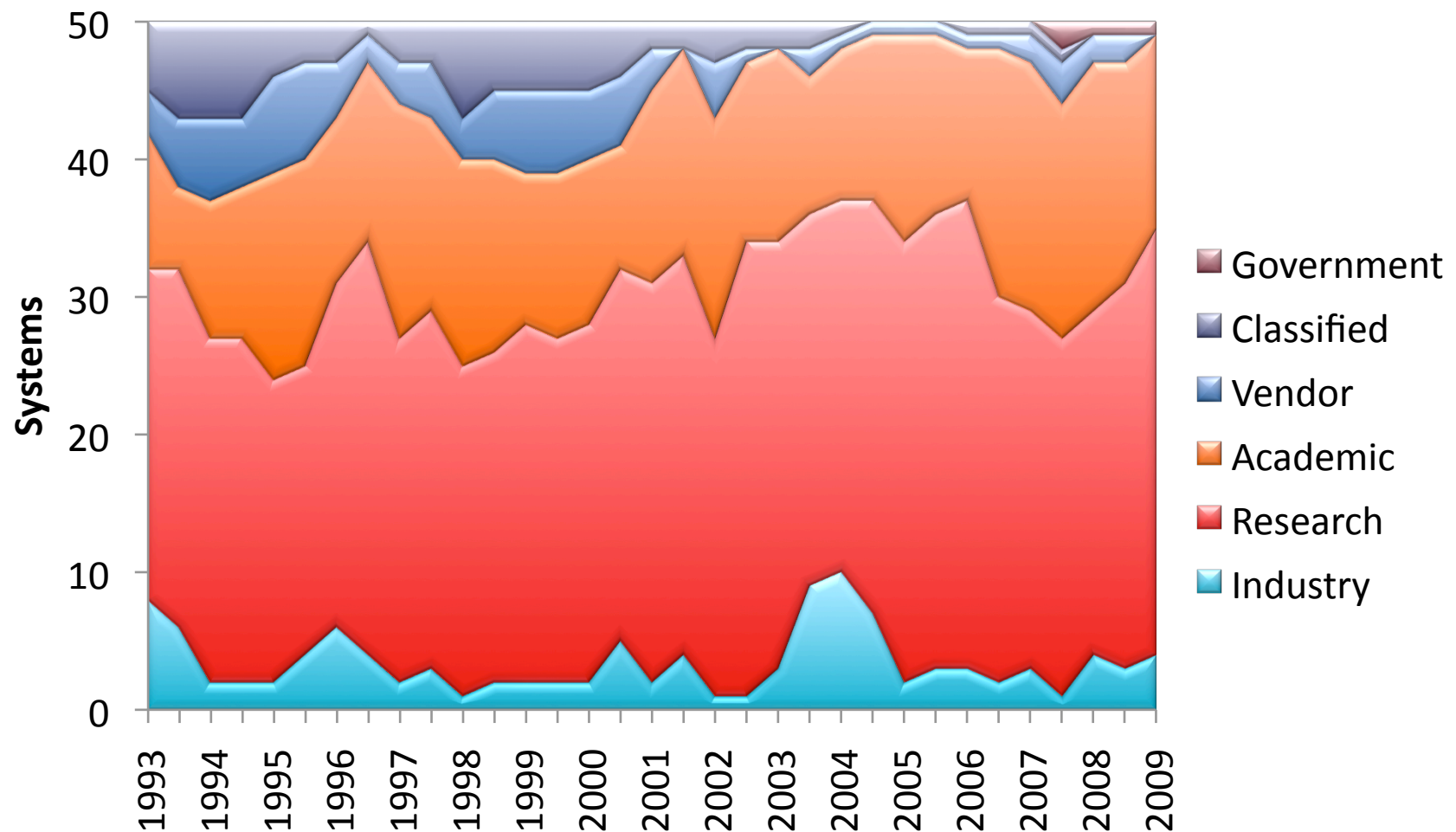
# Customer Segments



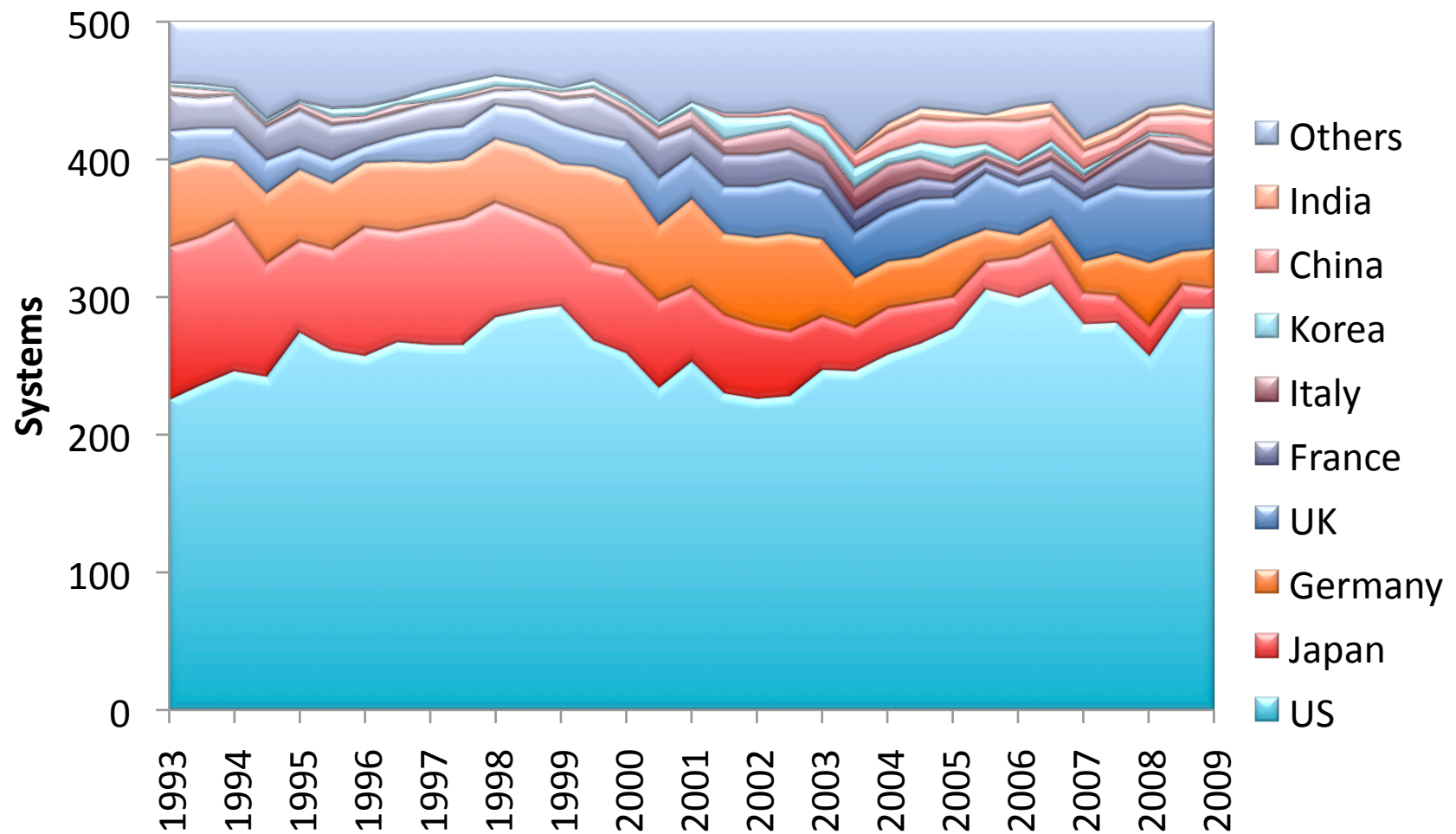
# Customer Segments



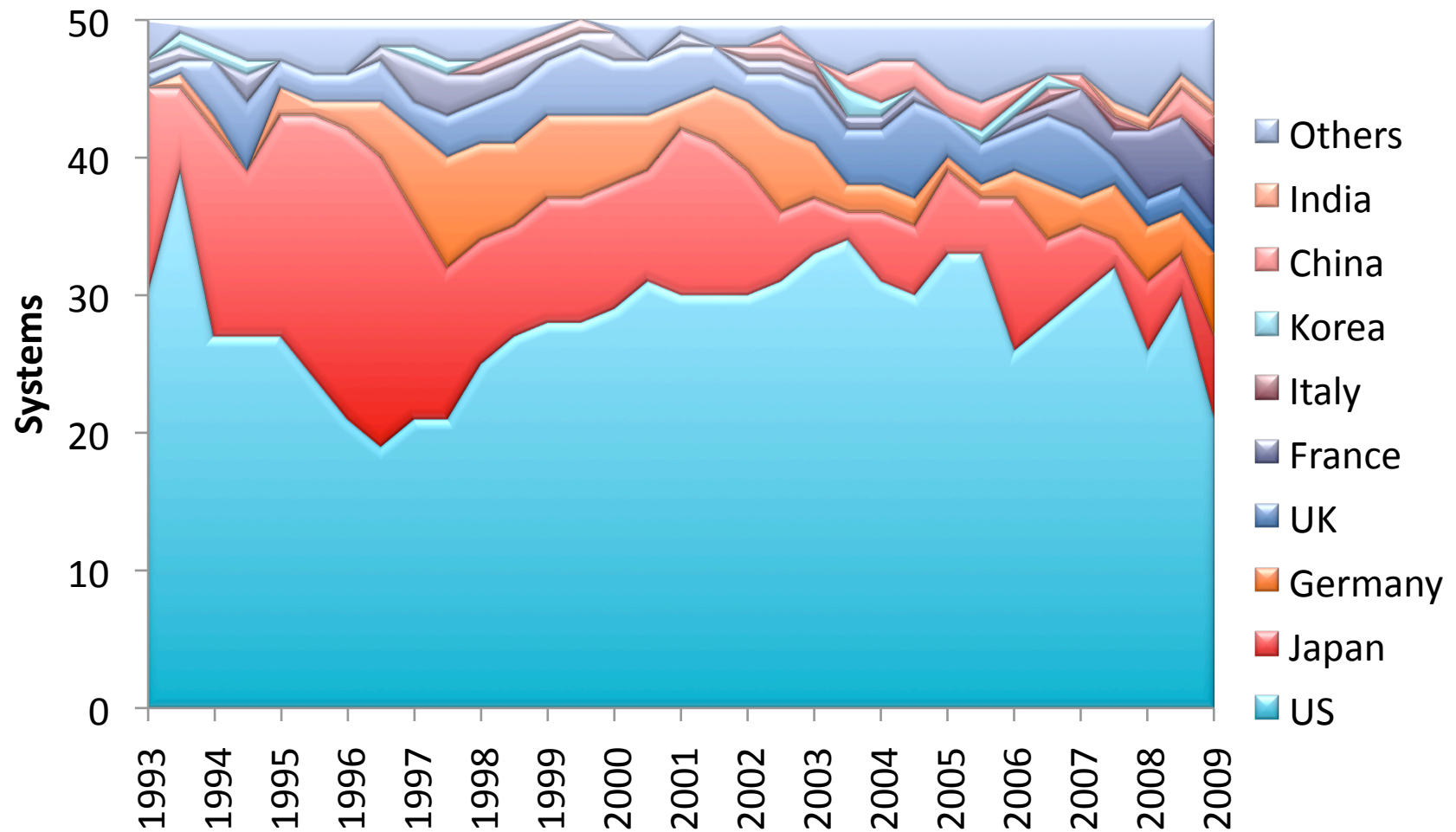
# Customer Segments (TOP50)



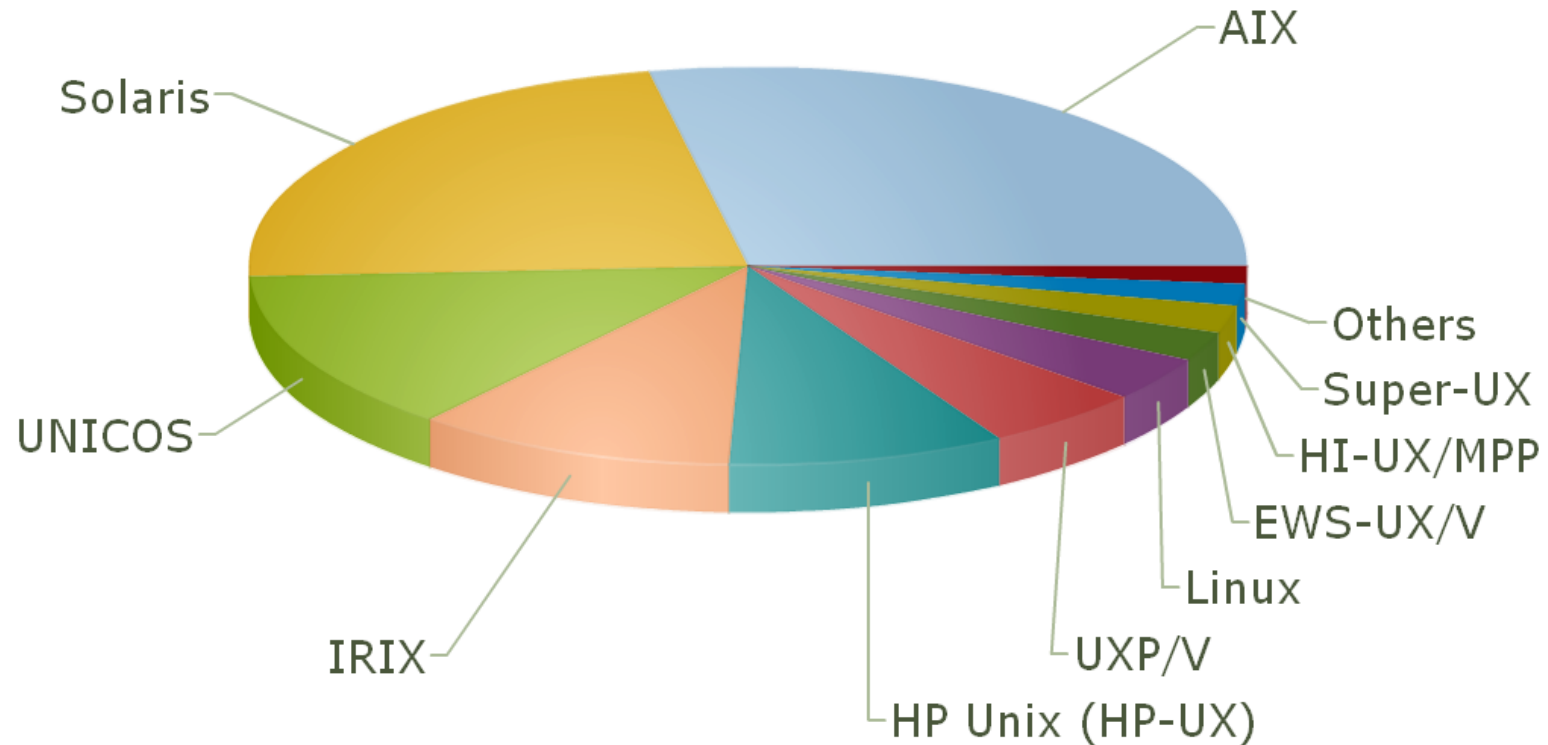
# Countries



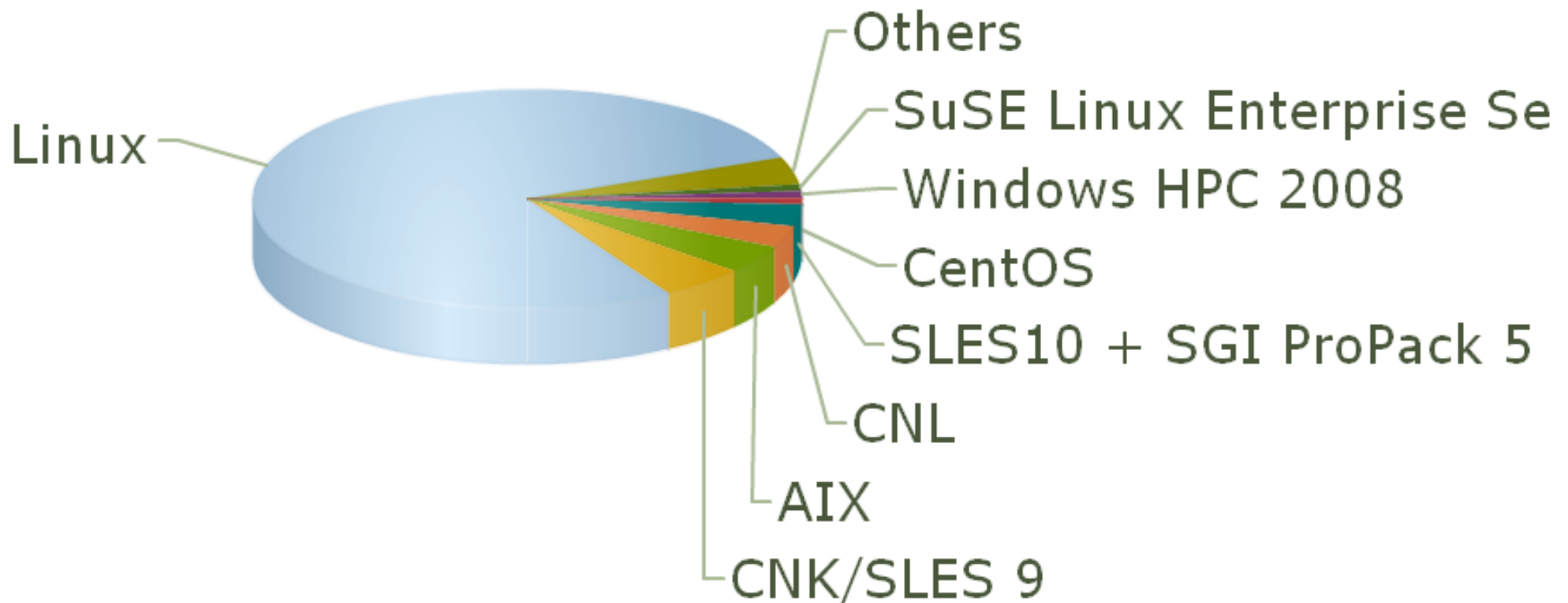
# Countries (TOP50)



# TOP500: OS Share Nov 1999



# TOP500: OS Share June 2009

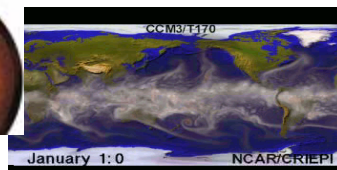
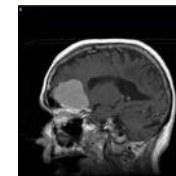




# Industrial Use of Supercomputers

- Of the 500 Fastest Supercomputer
  - Worldwide, Industrial Use is > 60%

- Aerospace
- Automotive
- Biology
- CFD
- Database
- Defense
- Digital Content Creation
- Digital Media
- Electronics
- Energy
- Environment
- Finance
- Gaming
- Geophysics
- Image Proc./Rendering
- Information Processing Service
- Information Service
- Life Science
- Media
- Medicine
- Pharmaceuticals
- Research
- Retail
- Semiconductor
- Telecomm
- Weather and Climate Research
- Weather Forecasting





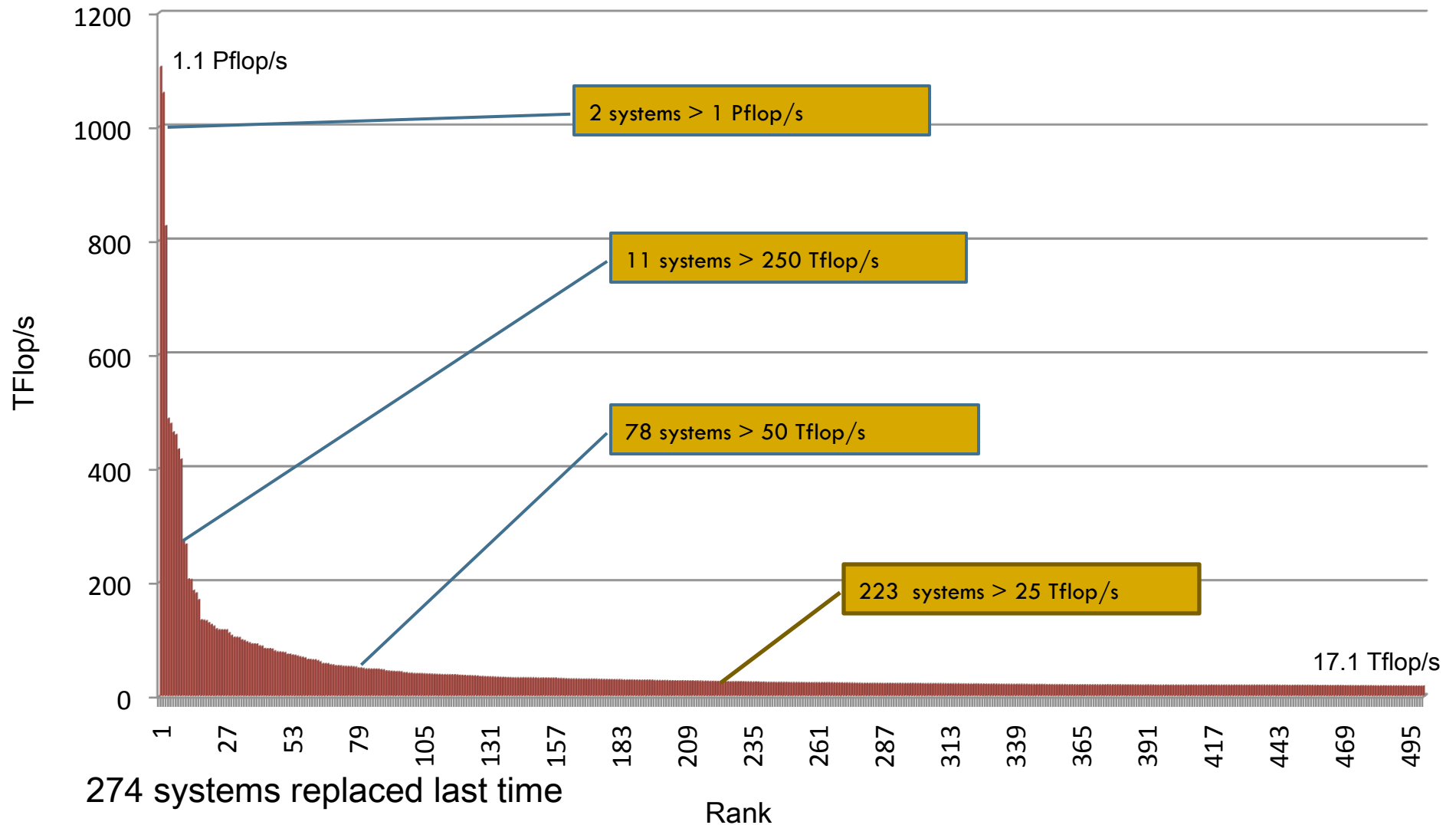
# 33<sup>rd</sup> List: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Tflops]	% of Peak
1	DOE / NNSA Los Alamos Nat Lab	Roadrunner / IBM BladeCenter QS22/LS21	USA	129,600	1,105	76
2	DOE / OS Oak Ridge Nat Lab	Jaguar / Cray Cray XT5 QC 2.3 GHz	USA	150,152	1,059	77
3	Forschungszentrum Juelich (FZJ)	Jugene / IBM Blue Gene/P Solution	Germany	294,912	825	82
4	NASA / Ames Research Center/NAS	Pleiades / SGI SGI Altix ICE 8200EX	USA	51,200	480	79
5	DOE / NNSA Lawrence Livermore NL	BlueGene/L IBM eServer Blue Gene Solution	USA	212,992	478	80
6	NSF NICS/U of Tennessee	Kraken / Cray Cray XT5 QC 2.3 GHz	USA	66,000	463	76
7	DOE / OS Argonne Nat Lab	Intrepid / IBM Blue Gene/P Solution	USA	163,840	458	82
8	NSF TACC/U. of Texas	Ranger / Sun SunBlade x6420	USA	62,976	433	75
9	DOE / NNSA Lawrence Livermore NL	Dawn / IBM Blue Gene/P Solution	USA	147,456	415	83
10	Forschungszentrum Juelich (FZJ)	JUROPA /Sun - Bull SA NovaScale /Sun Blade	Germany	26,304	274	89

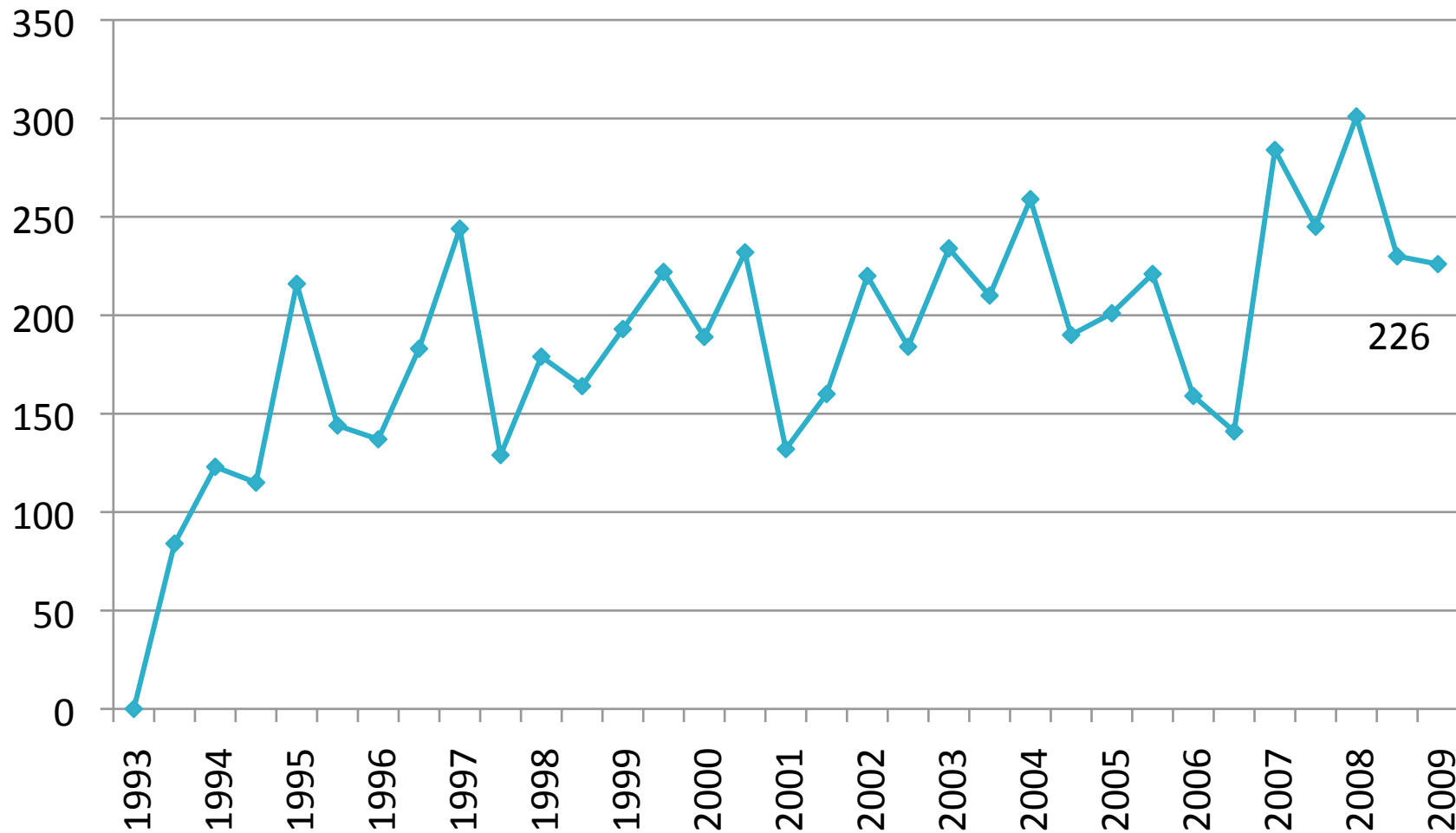
# 33<sup>rd</sup> List: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Tflops]	% of Peak	Power [MW]	Flops/ Watt
1	DOE / NNSA Los Alamos Nat Lab	Roadrunner / IBM BladeCenter QS22/LS21	USA	129,600	1,105	76	2.48	446
2	DOE / OS Oak Ridge Nat Lab	Jaguar / Cray Cray XT5 QC 2.3 GHz	USA	150,152	1,059	77	6.95	151
3	Forschungszentrum Juelich (FZJ)	Jugene / IBM Blue Gene/P Solution	Germany	294,912	825	82	2.26	365
4	NASA / Ames Research Center/NAS	Pleiades / SGI SGI Altix ICE 8200EX	USA	51,200	480	79	2.09	230
5	DOE / NNSA Lawrence Livermore NL	BlueGene/L IBM eServer Blue Gene Solution	USA	212,992	478	80	2.32	206
6	NSF NICS/U of Tennessee	Kraken / Cray Cray XT5 QC 2.3 GHz	USA	66,000	463	76		
7	DOE / OS Argonne Nat Lab	Intrepid / IBM Blue Gene/P Solution	USA	163,840	458	82	1.26	363
8	NSF TACC/U. of Texas	Ranger / Sun SunBlade x6420	USA	62,976	433	75	2.0	217
9	DOE / NNSA Lawrence Livermore NL	Dawn / IBM Blue Gene/P Solution	USA	147,456	415	83	1.13	367
10	Forschungszentrum Juelich (FZJ)	JUROPA /Sun - Bull SA NovaScale /Sun Blade	Germany	26,304	274	89	1.54	178

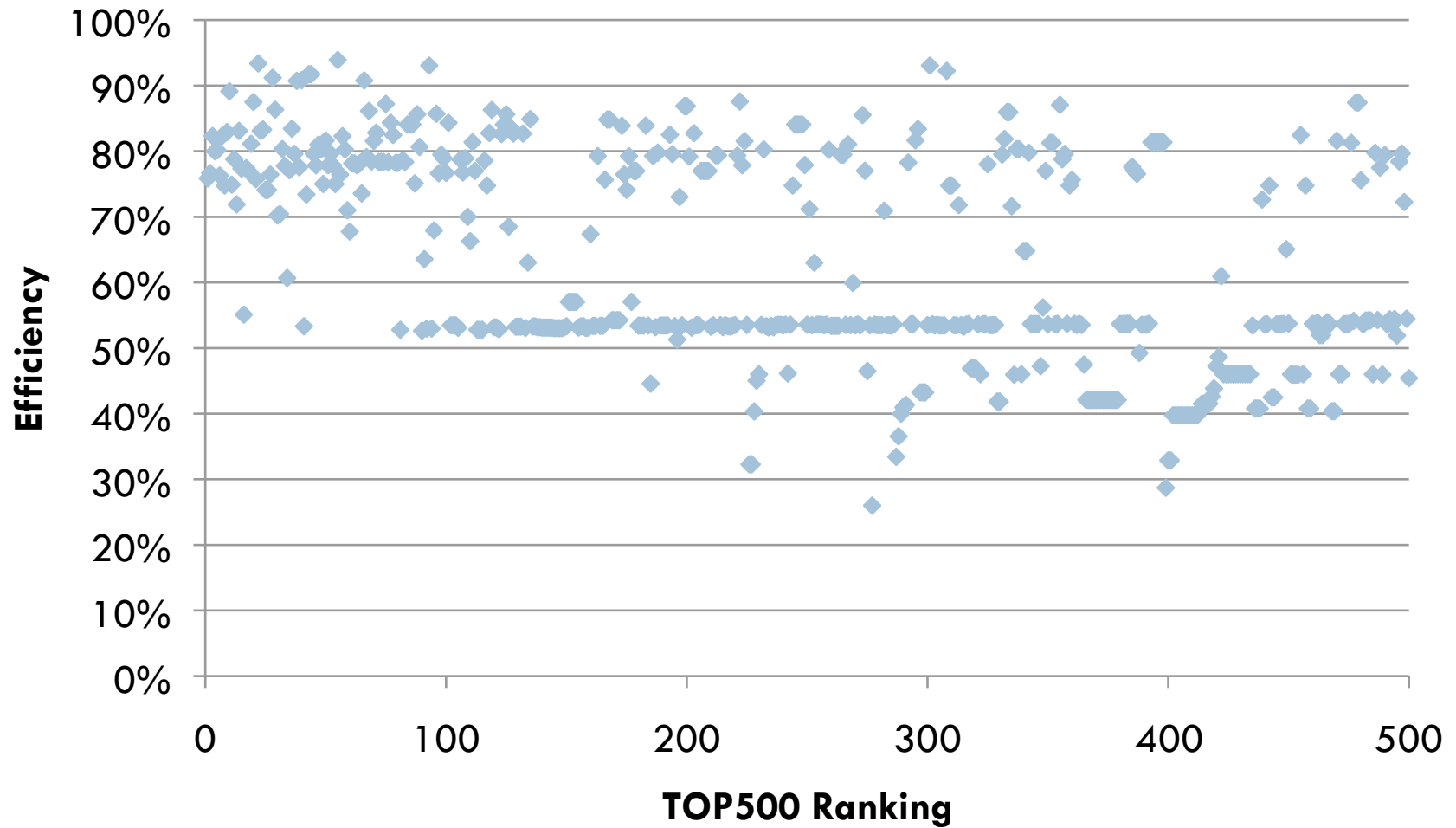
# Distribution of the Top500



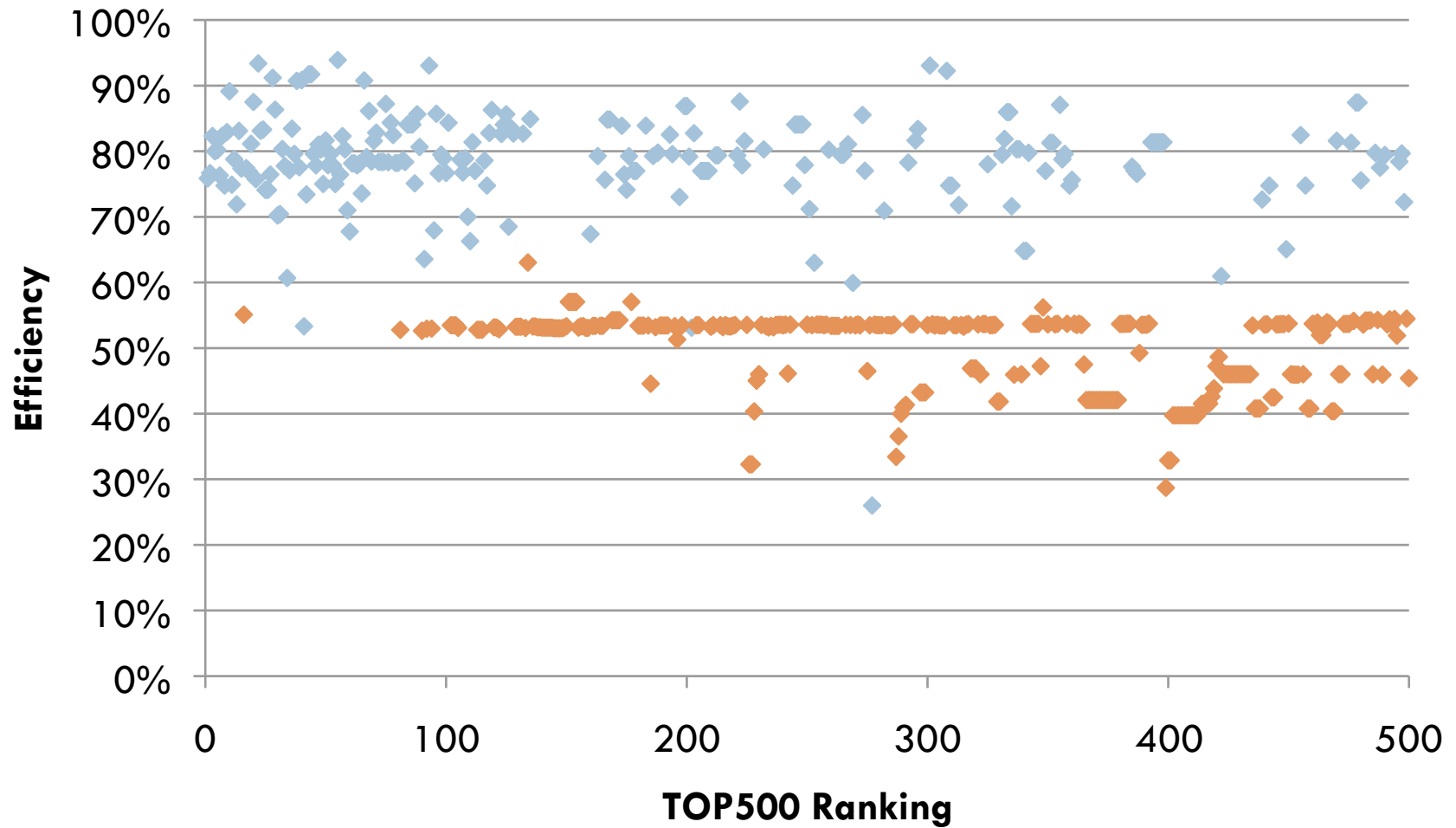
# Replacement Rate



# Linpack Efficiency



# Linpack Efficiency

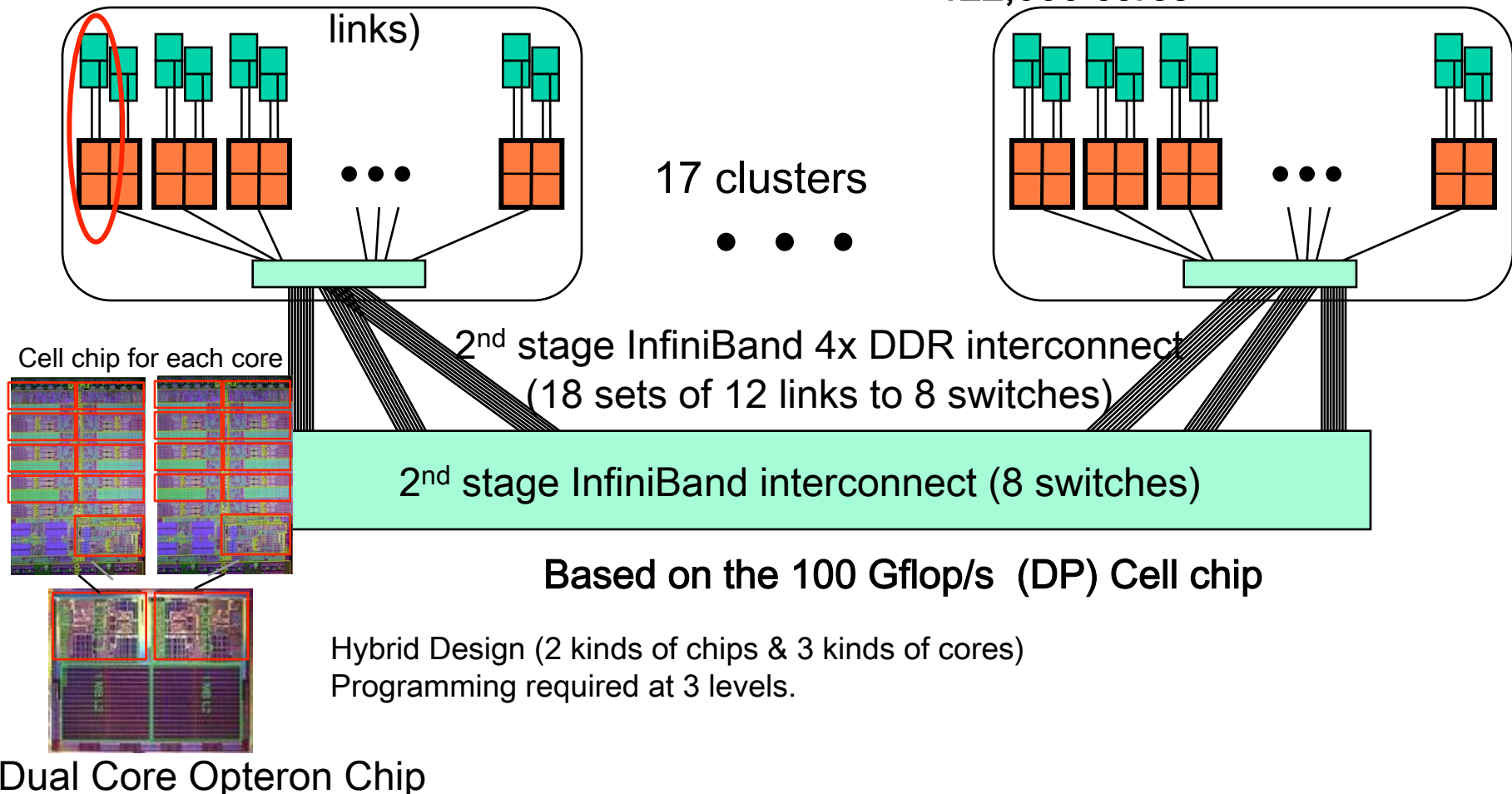


# LANL Roadrunner

## A Petascale System in 2008

“Connected Unit” cluster  
192 Opteron nodes  
(180 w/ 2 dual-Cell blades  
connected w/ 4 PCIe x8

≈ 13,000 Cell HPC chips  
≈ 1.33 PetaFlop/s (from Cell)  
≈ 7,000 dual-core Opterons  
≈ 122,000 cores





# ORNL's Newest System Jaguar XT5



Jaguar	Total	XT5	XT4
Peak Performance	1,645	1,382	263
AMD Opteron Cores	181,504	150,176	31,328
System Memory (TB)	362	300	62
Disk Bandwidth (GB/s)	284	240	44
Disk Space (TB)	10,750	10,000	750
Interconnect Bandwidth (TB/s)	532	374	157

Will be upgraded this year to a 2 Pflop/s system with > 224K AMD Istanbul Cores.



U.S. DEPARTMENT OF  
**ENERGY**

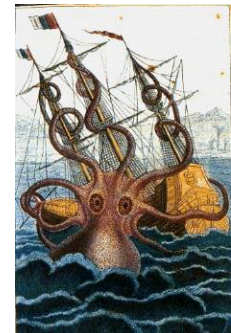
Office of  
Science





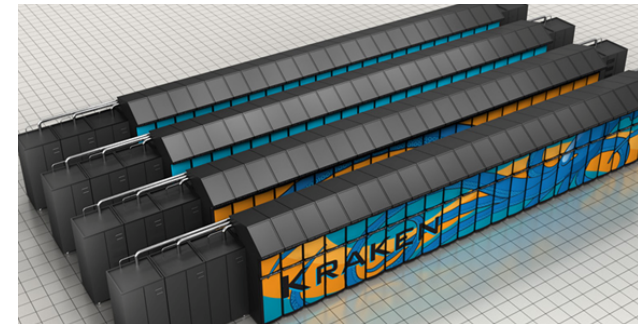
# 's HPC System

- University of Tennessee's National Institute for Computational Sciences
- Housed at ORNL, operated for the NSF, named Kraken



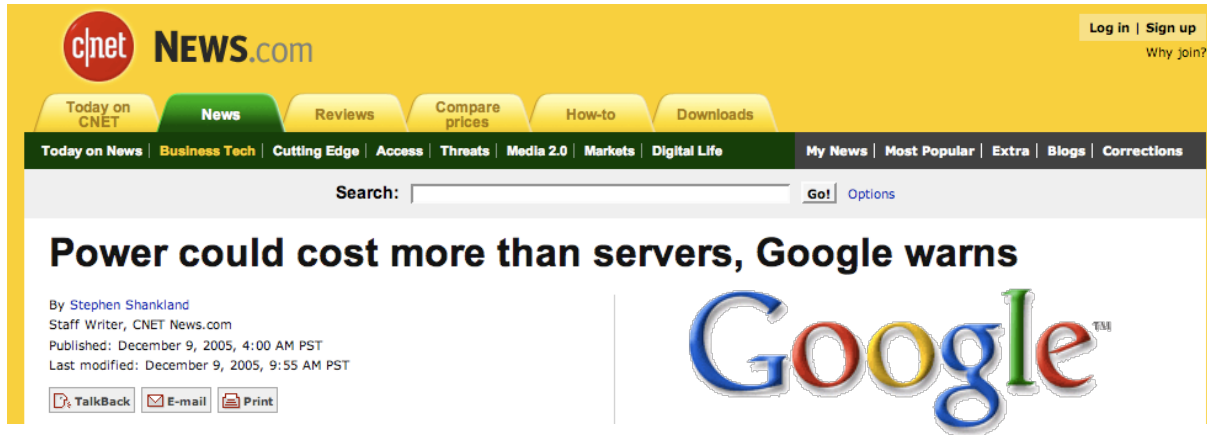
Today:

- Cray XT5 (608 TF) + Cray XT4 (167 TF)
  - XT5: 16,512 sockets, 66,048 cores
  - XT4: 4,512 sockets, 18,048 cores
- Number 6 on the Top500



Later 2009: upgrading to 1 Pflop/s

# Power is an Industry Wide Problem



- ◆ **Google facilities**
  - **leveraging hydroelectric power**
  - **old aluminum plants**

**The New York Times** “Hiding in Plain Sight, Google Seeks More Power”,  
by John Markoff, June 14, 2006

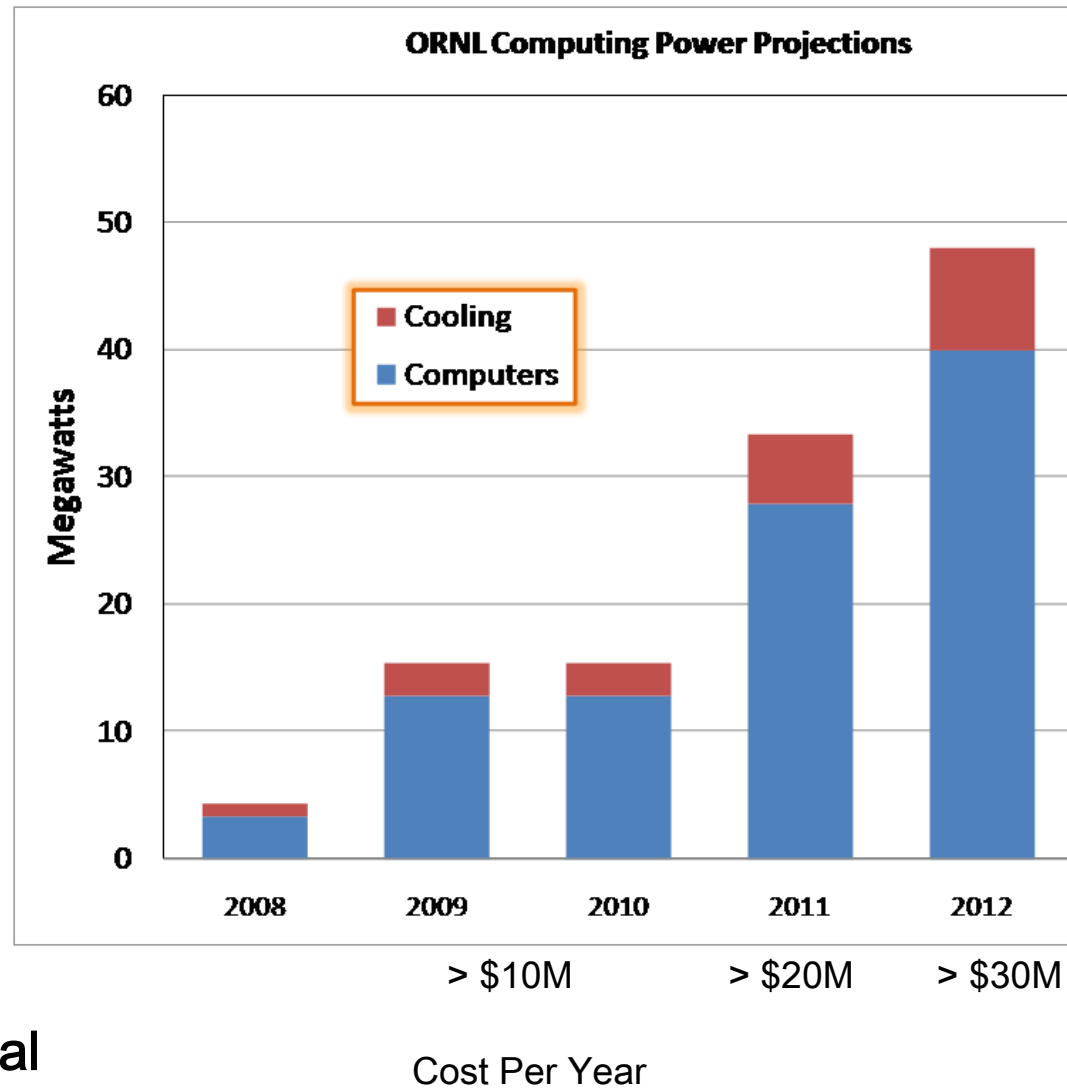


Microsoft and Yahoo are building big data centers upstream in Wenatchee and Quincy, Wash.  
– To keep up with Google, which means they need cheap electricity and readily accessible data networking

Microsoft Quincy, Wash.  
470,000 Sq Ft, 47MW!

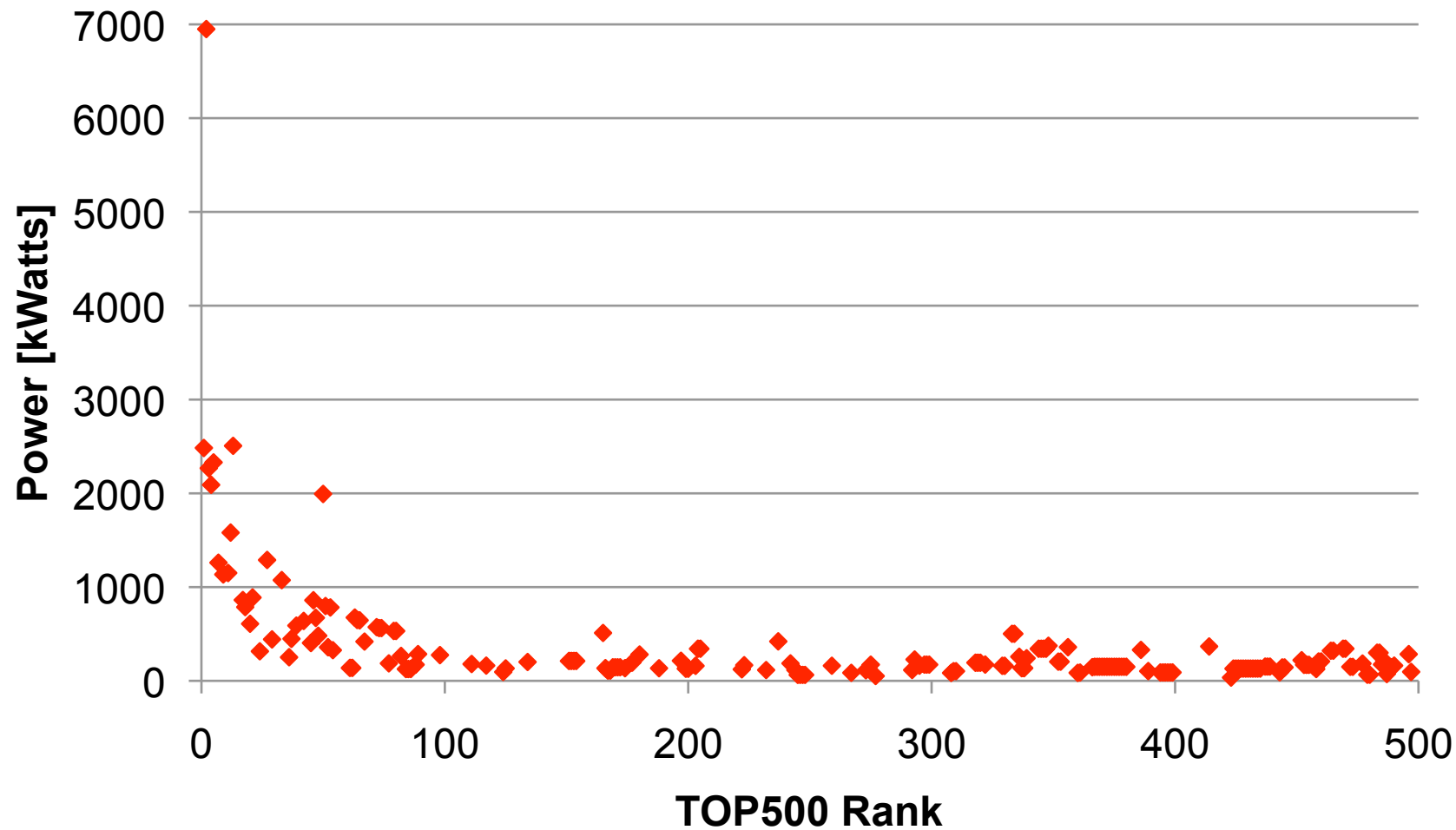
# ORNL/UTK Computer Power Cost Projections 2008-2012

- Over the next 5 years ORNL/UTK will deploy 2 large Petascale systems
- Using 15 MW today
- By 2012 close to 50MW!!
- Power costs close to \$10M today.
- Cost estimates based on \$0.07 per Kwh

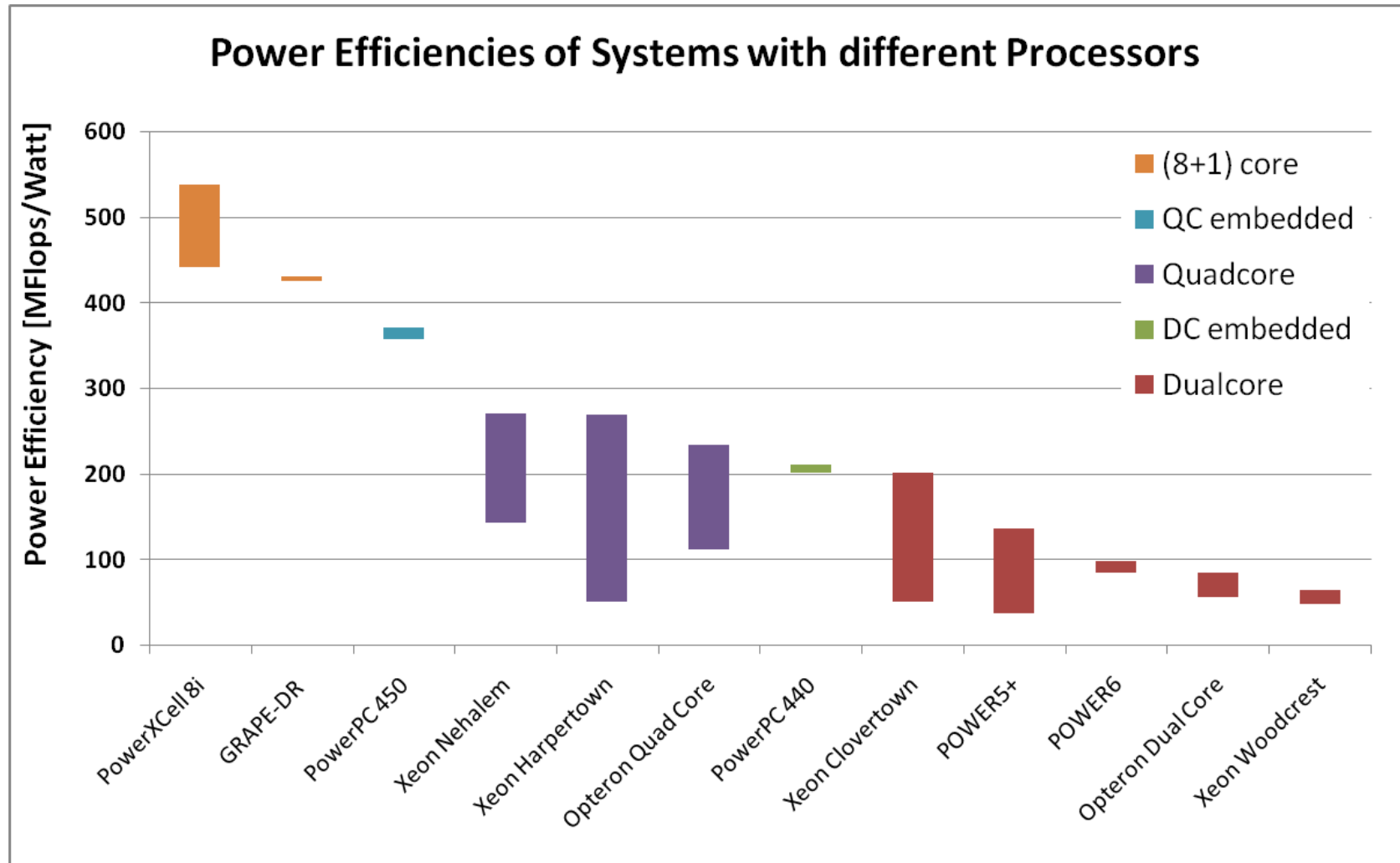


Power becomes the architectural driver for future large systems

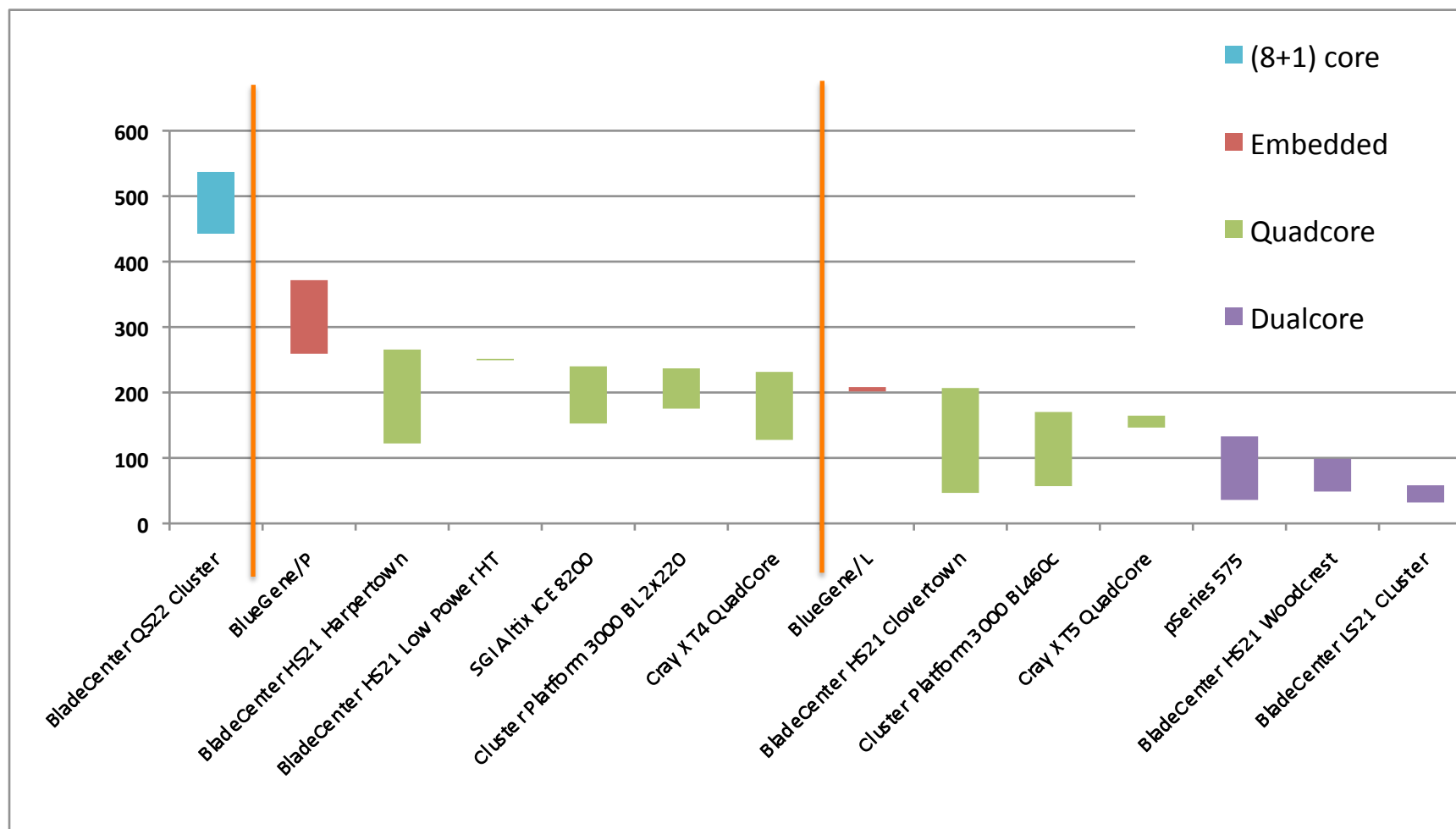
# Absolute Power Levels



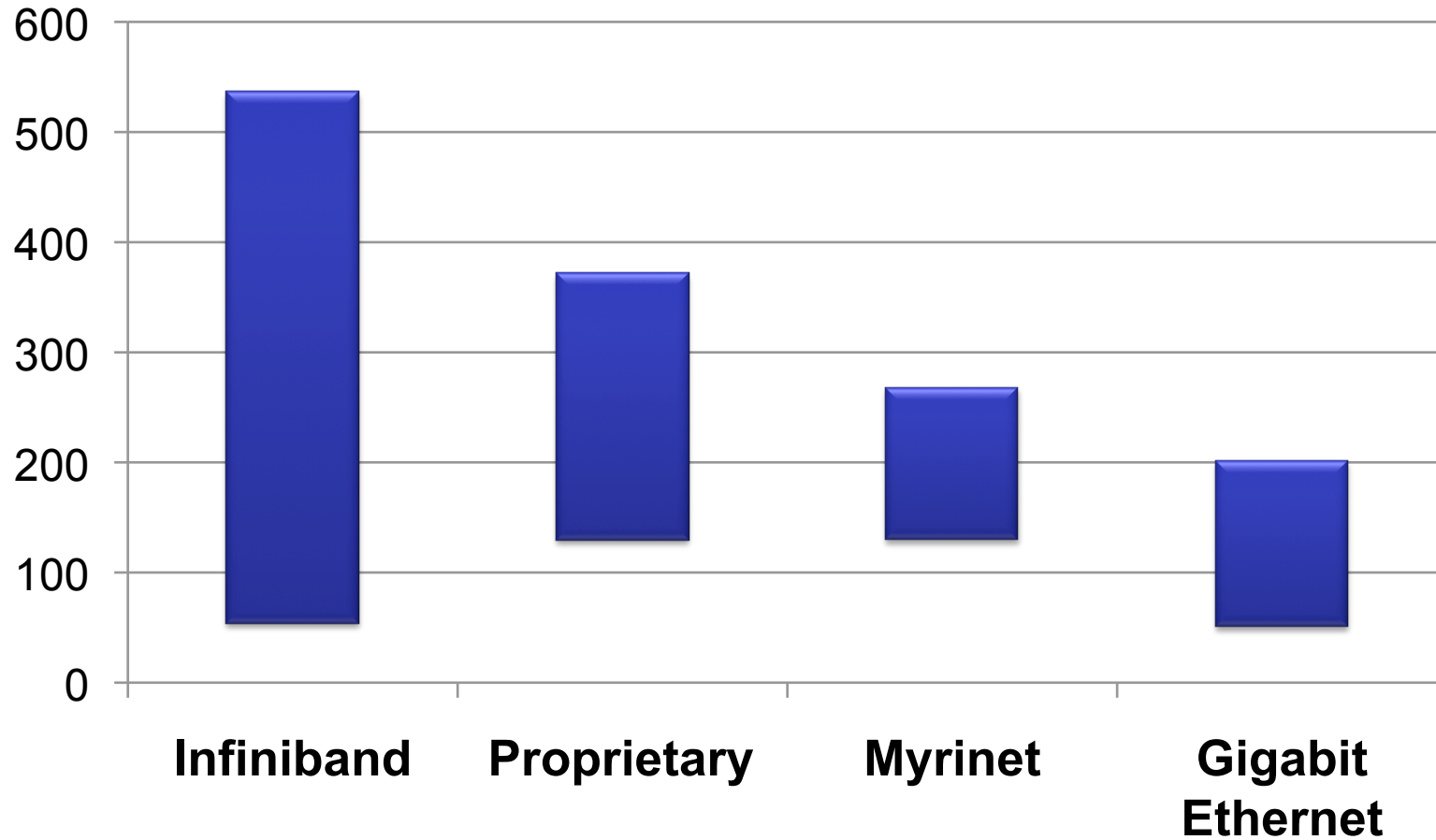
# Power Efficiency related to Processors



# Power Efficiencies of different Systems

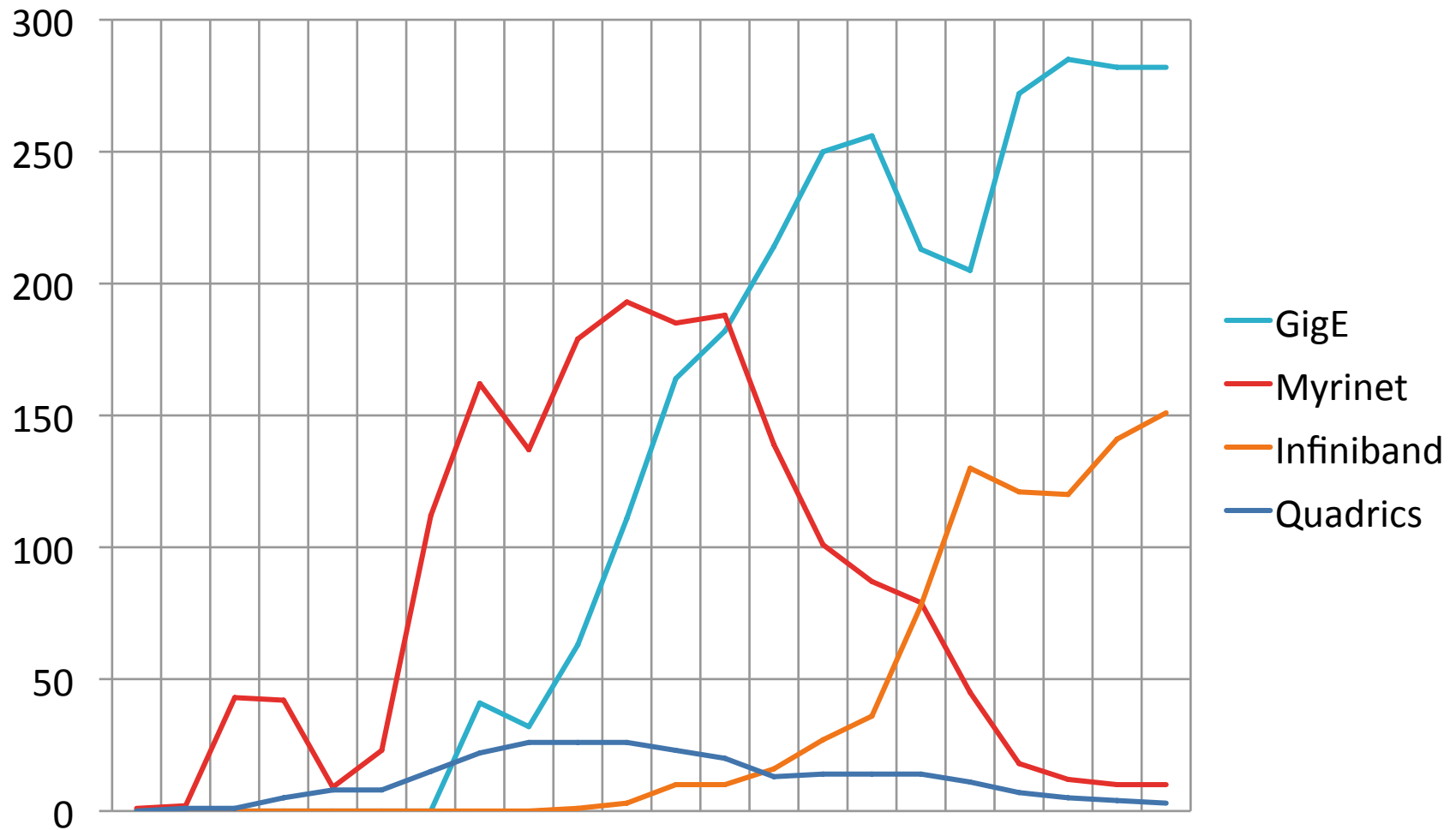


# Power Efficiency related to Interconnects

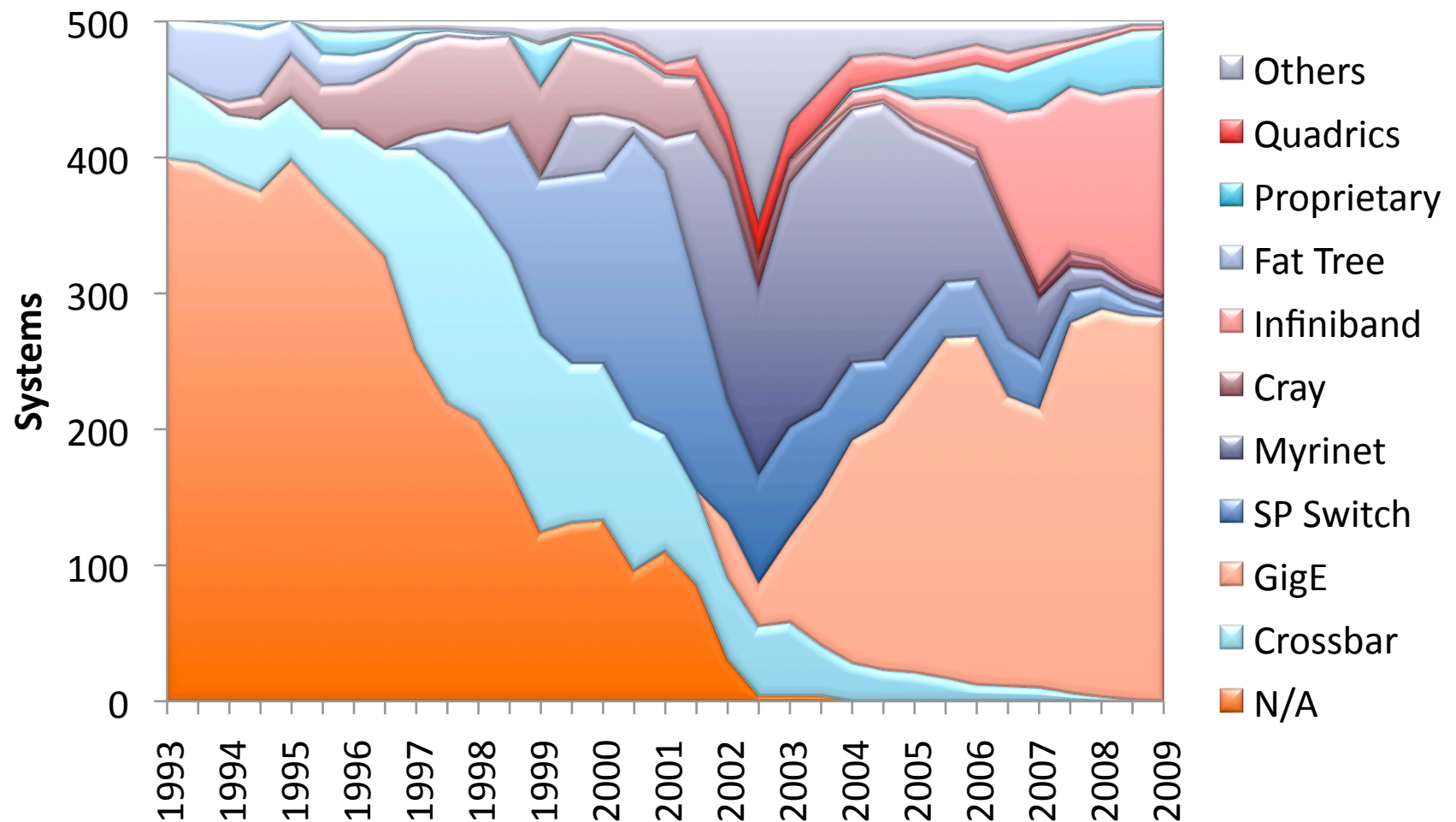




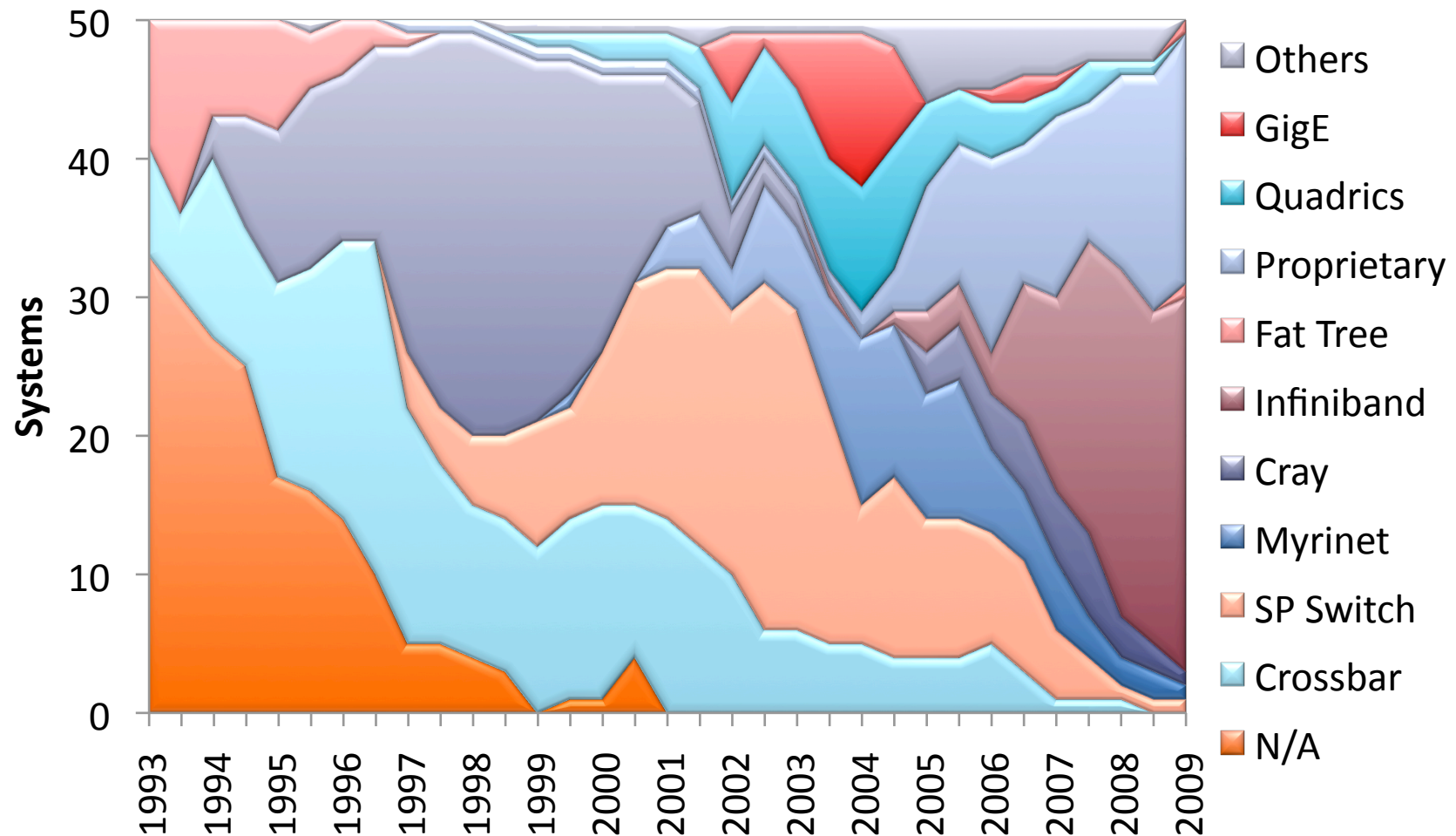
# Cluster Interconnects



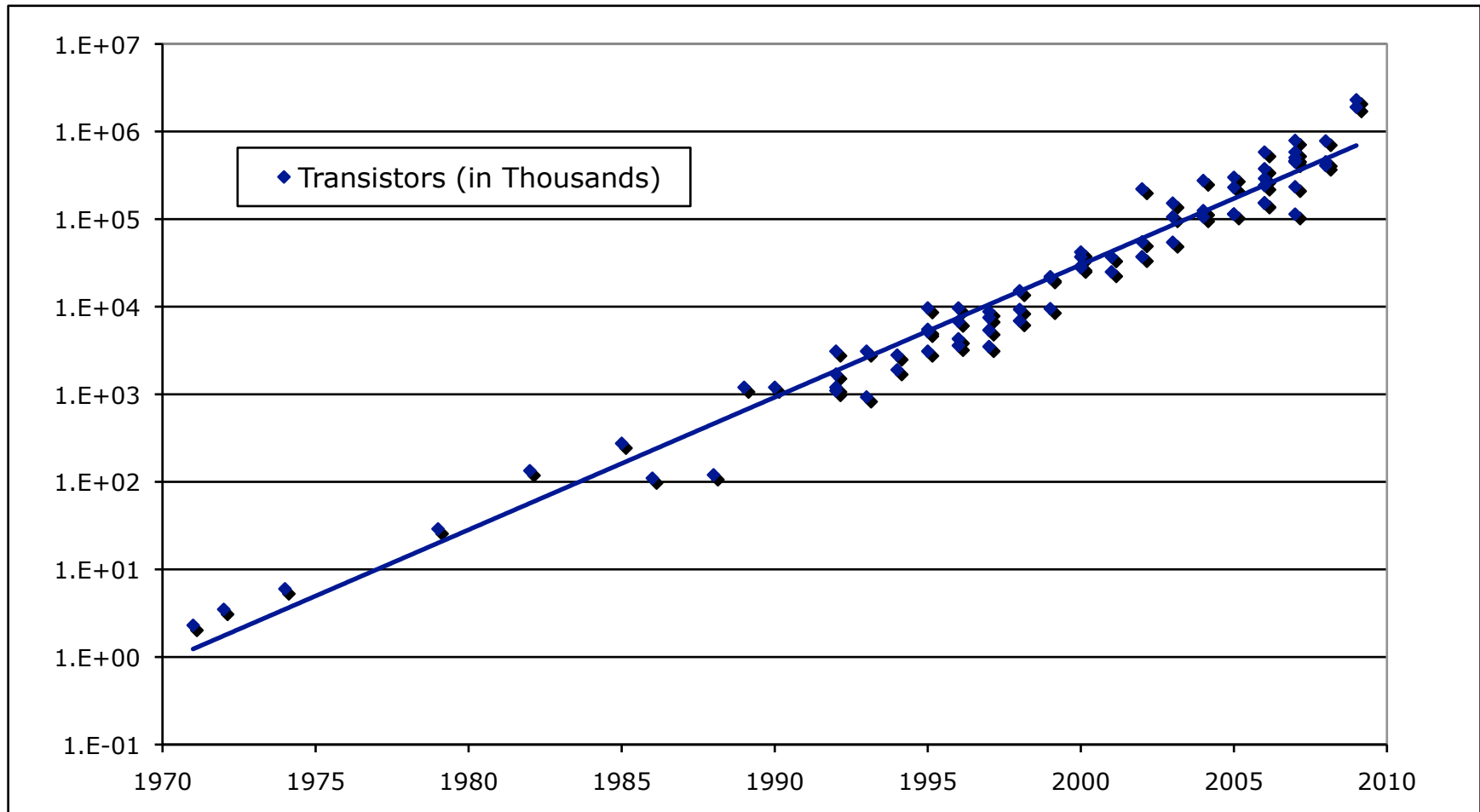
# Interconnect Family



# Interconnect Family (TOP50)



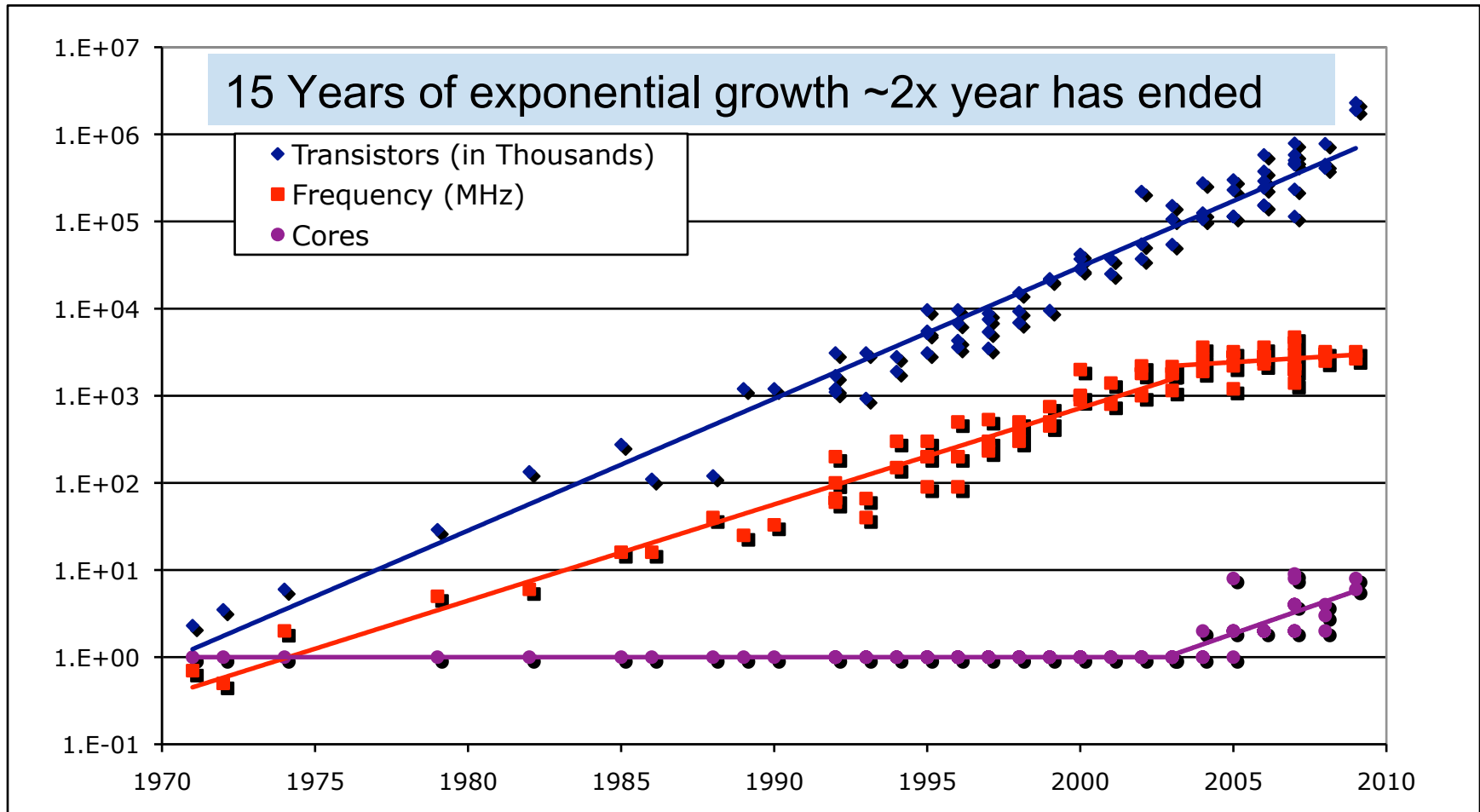
# Moore's Law is Alive and Well



Data from Kunle Olukotun, Lance Hammond, Herb Sutter,  
Burton Smith, Chris Batten, and Krste Asanović

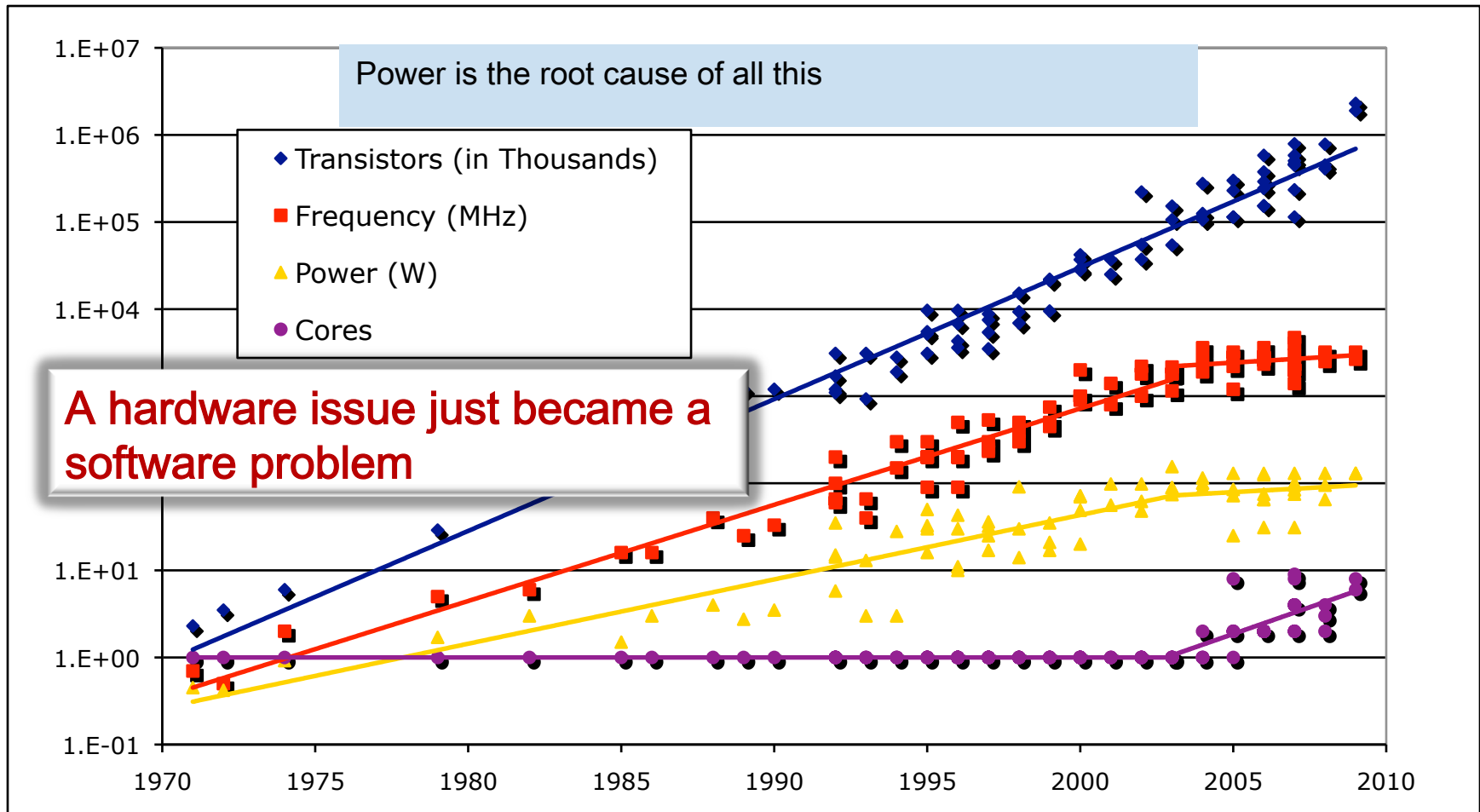


# But Clock Frequency Scaling Replaced by Scaling Cores / Chip



Data from Kunle Olukotun, Lance Hammond, Herb Sutter,  
Burton Smith, Chris Batten, and Krste Asanović

# Performance Has Also Slowed, Along with Power



Data from Kunle Olukotun, Lance Hammond, Herb Sutter,  
Burton Smith, Chris Batten, and Krste Asanović

# Moore's Law Reinterpreted

---

- Number of cores per chip doubles every 2 year, while clock speed decreases (not increases).
  - Need to deal with systems with millions of concurrent threads
    - Future generation will have billions of threads!
  - Need to be able to easily replace inter-chip parallelism with intro-chip parallelism
- Number of threads of execution doubles every 2 year

# Power Cost of Frequency

- Power  $\propto$  Voltage<sup>2</sup> x Frequency (V<sup>2</sup>F)
- Frequency  $\propto$  Voltage
- Power  $\propto$  Frequency<sup>3</sup>

	Cores	V	Freq	Perf	Power	PE (Bops/watt)
Superscalar	1	1	1	1	1	1
"New" Superscalar	1X	1.5X	1.5X	1.5X	3.3X	0.45X



# Power Cost of Frequency

- Power  $\propto$  Voltage<sup>2</sup> x Frequency (V<sup>2</sup>F)
- Frequency  $\propto$  Voltage
- Power  $\propto$  Frequency<sup>3</sup>

	Cores	V	Freq	Perf	Power	PE (Bops/watt)
Superscalar	1	1	1	1	1	1
"New" Superscalar	1X	1.5X	1.5X	1.5X	3.3X	0.45X
Multicore	2X	0.75X	0.75X	1.5X	0.8X	1.88X

(Bigger # is better)

50% more performance with 20% less power

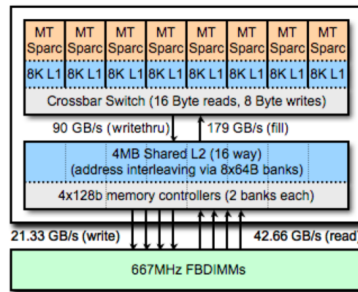
Preferable to use multiple slower devices, than one superfast device



# Today's Multicores

99% of Top500 Systems Are Based on Multicore

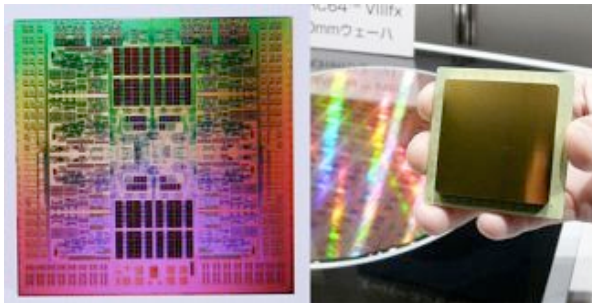
282 use Quad-Core  
204 use Dual-Core  
3 use Nona-core



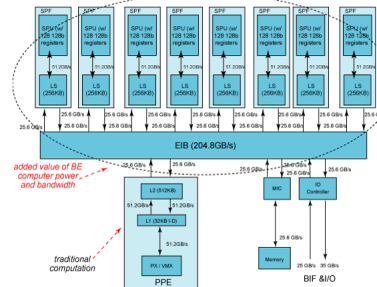
Sun Niagara2 (8 cores)



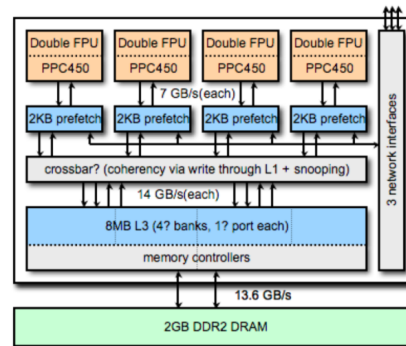
IBM Power 7 (8 cores)



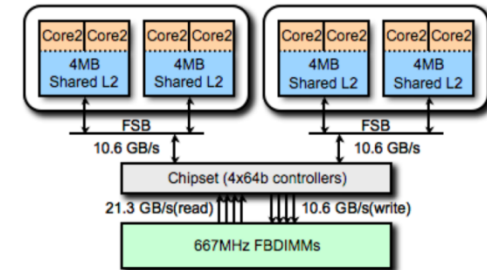
Fujitsu Venus (8 cores)



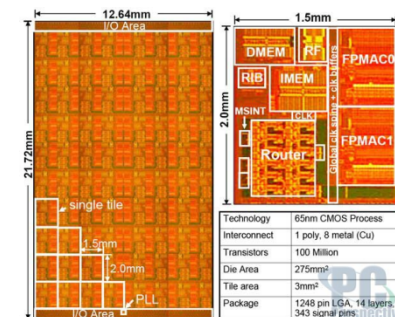
IBM Cell (9 cores)



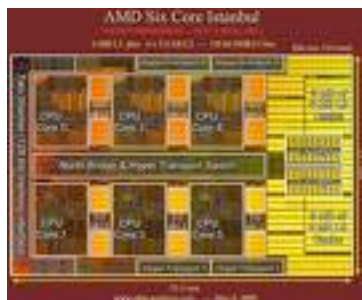
IBM BG/P (4 cores)



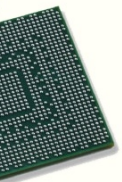
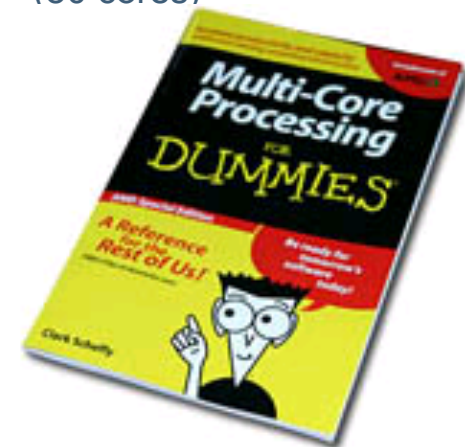
Intel Cloverton (4 cores)



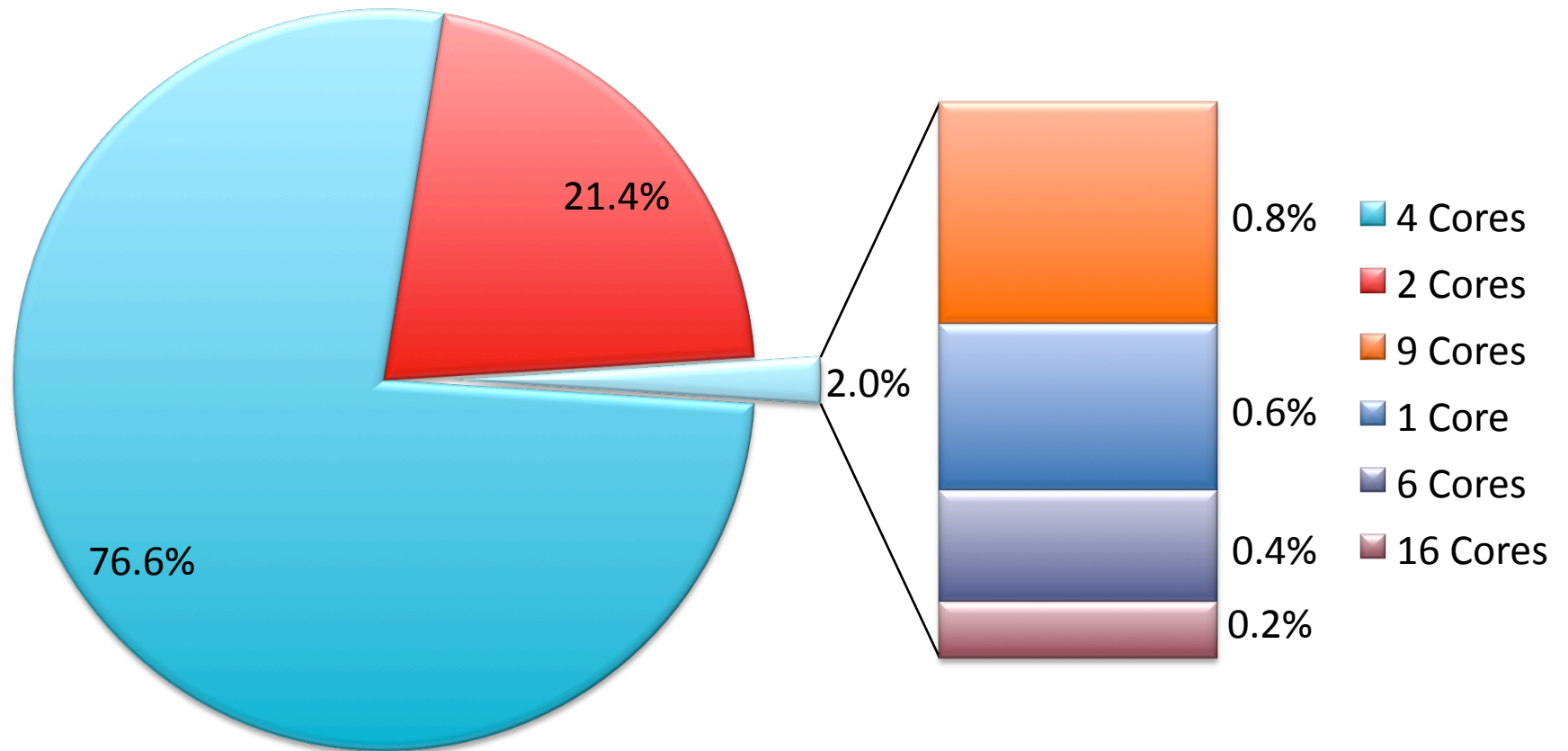
Intel Polarix [experimental]  
(80 cores)



AMD Istanbul (6 cores)

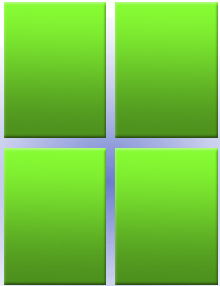


# Cores per Socket

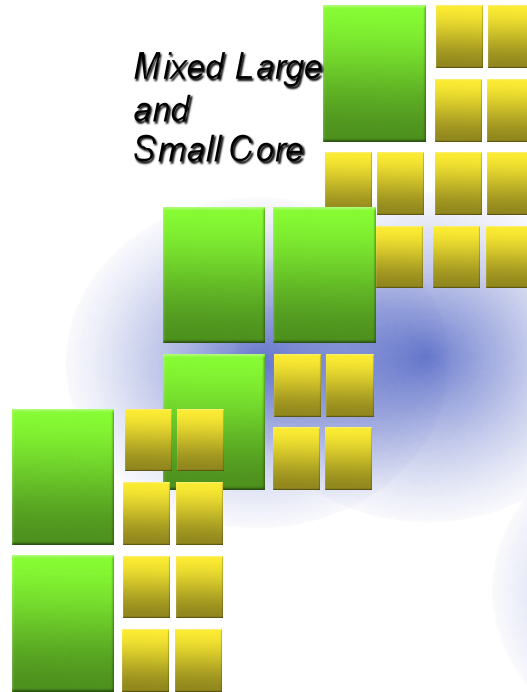


# What's Next?

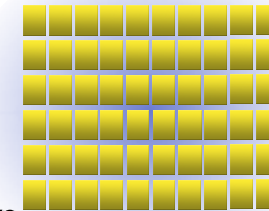
All Large Core



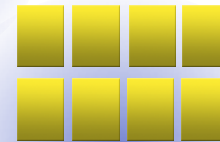
Mixed Large and Small Core



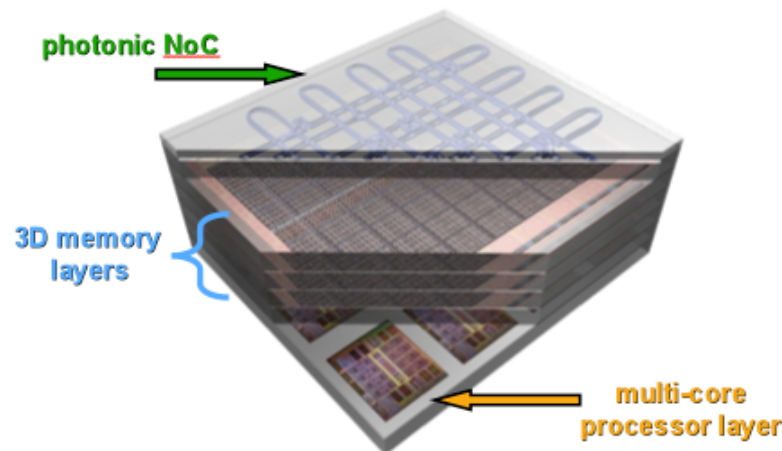
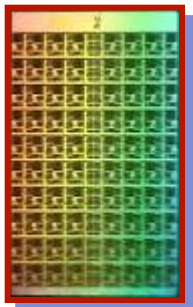
Many Small Cores



All Small Core



Many Floating-Point Cores



+ 3D Stacked Memory

## Different Classes of Chips

- Home
- Games / Graphics
- Business
- Scientific



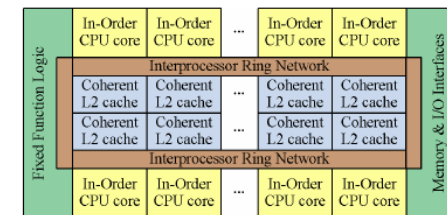
# Commodity

---

- **Moore's "Law" favored consumer commodities**
  - Economics drove enormous improvements
  - Specialized processors and mainframes faltered
  - Custom HPC hardware largely disappeared
  - Hard to compete against 50%/year improvement
- **Implications**
  - Consumer product space defines outcomes
  - It does not always go where we hope or expect
  - Research environments track commercial trends
  - Driven by market economics
  - Think about processors, clusters, commodity storage

# Future Computer Systems

- Most likely be a hybrid design
- Think standard multicore chips and accelerator (GPUs)
- Today accelerators are attached
- Next generation more integrated
- Intel's Larrabee in 2010
  - 8,16,32,or 64 x86 cores
- AMD's Fusion in 2011
  - Multicore with embedded graphics ATI
- Nvidia's plans?



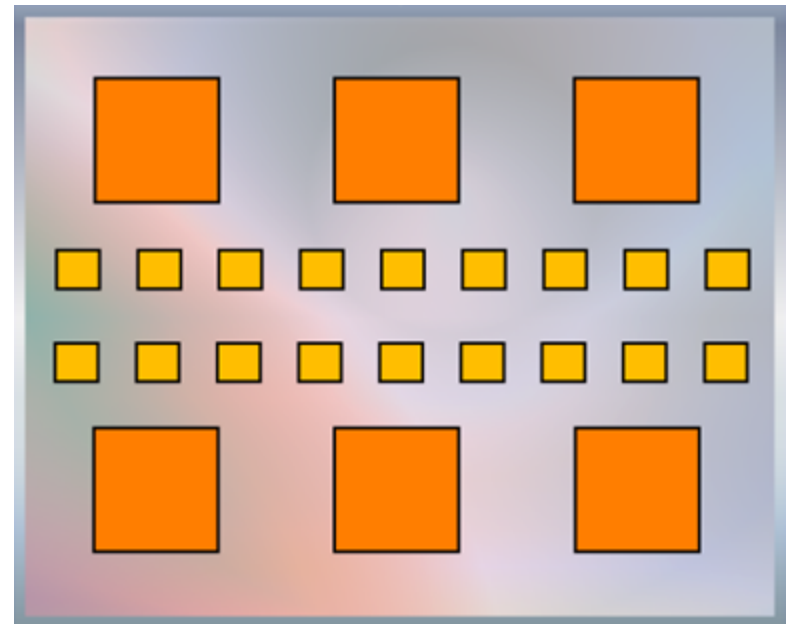
Intel Larrabee



# Architecture of Interest

---

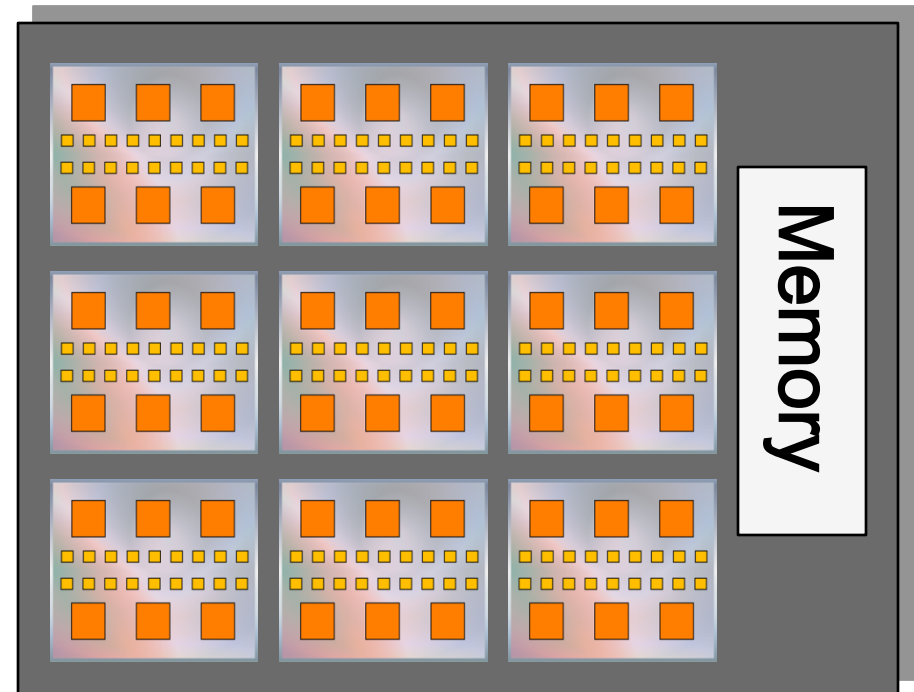
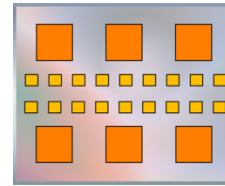
- **Manycore chip**
- **Composed of hybrid cores**
  - **Some general purpose**
  - **Some graphics**
  - **Some floating point**



# Architecture of Interest

---

- Board composed of multiple chips sharing memory

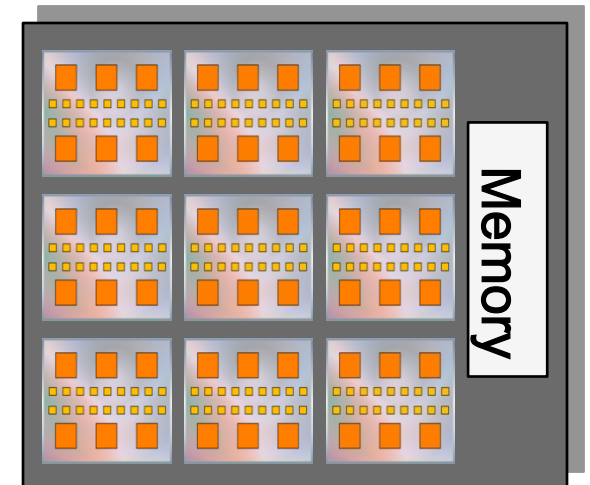
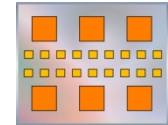
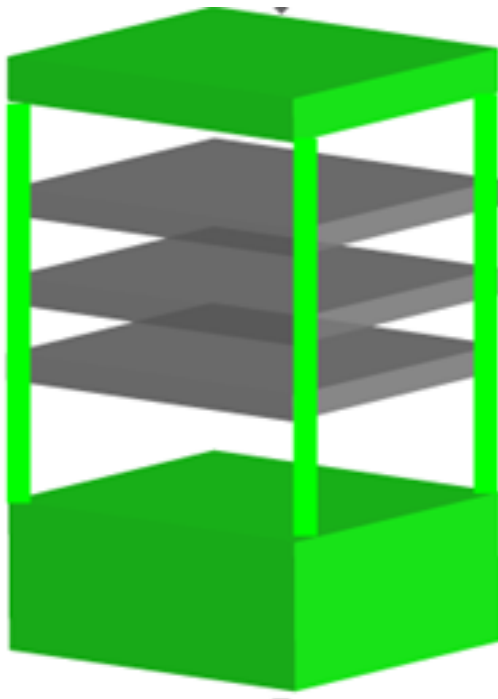




# Architecture of Interest

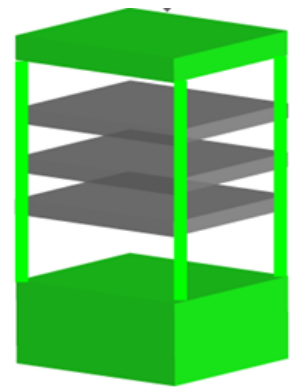
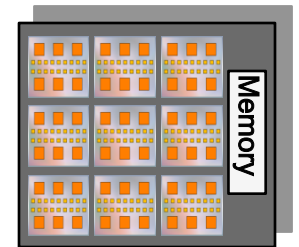
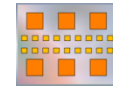
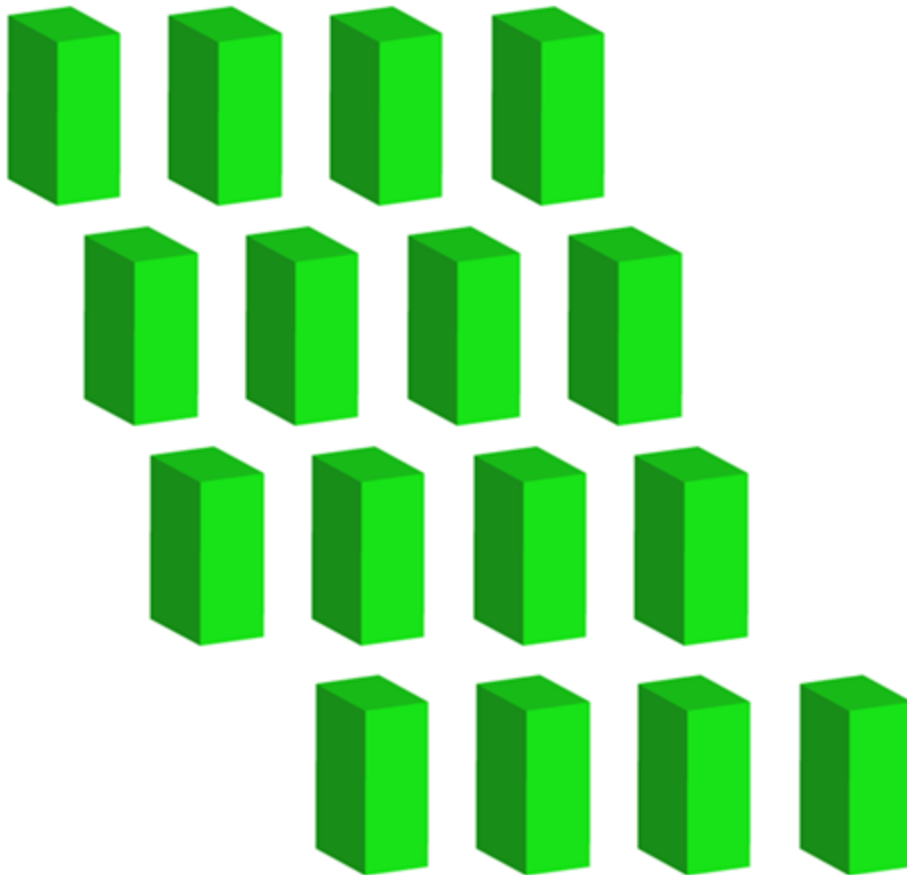
---

- Rack composed of multiple boards



# Architecture of Interest

- A room full of these racks



- Think millions of cores

# Major Changes to Software

---

- **Must rethink the design of our software**
  - **Another disruptive technology**
    - Similar to what happened with cluster computing and message passing
  - **Rethink and rewrite the applications, algorithms, and software**
- **Numerical libraries for example will change**
  - **For example, both LAPACK and ScaLAPACK will undergo major changes to accommodate this**



# Quasi Mainstream Programming Models

---

- C, Fortran, C++ and MPI
- OpenMP, pthreads
- (CUDA, RapidMind, Cn) → OpenCL
- PGAS (UPC, CAF, Titanium)
- HPCS Languages (Chapel, Fortress, X10)
- HPC Research Languages and Runtime
- HLL (Parallel Matlab, Grid Mathematica, etc.)

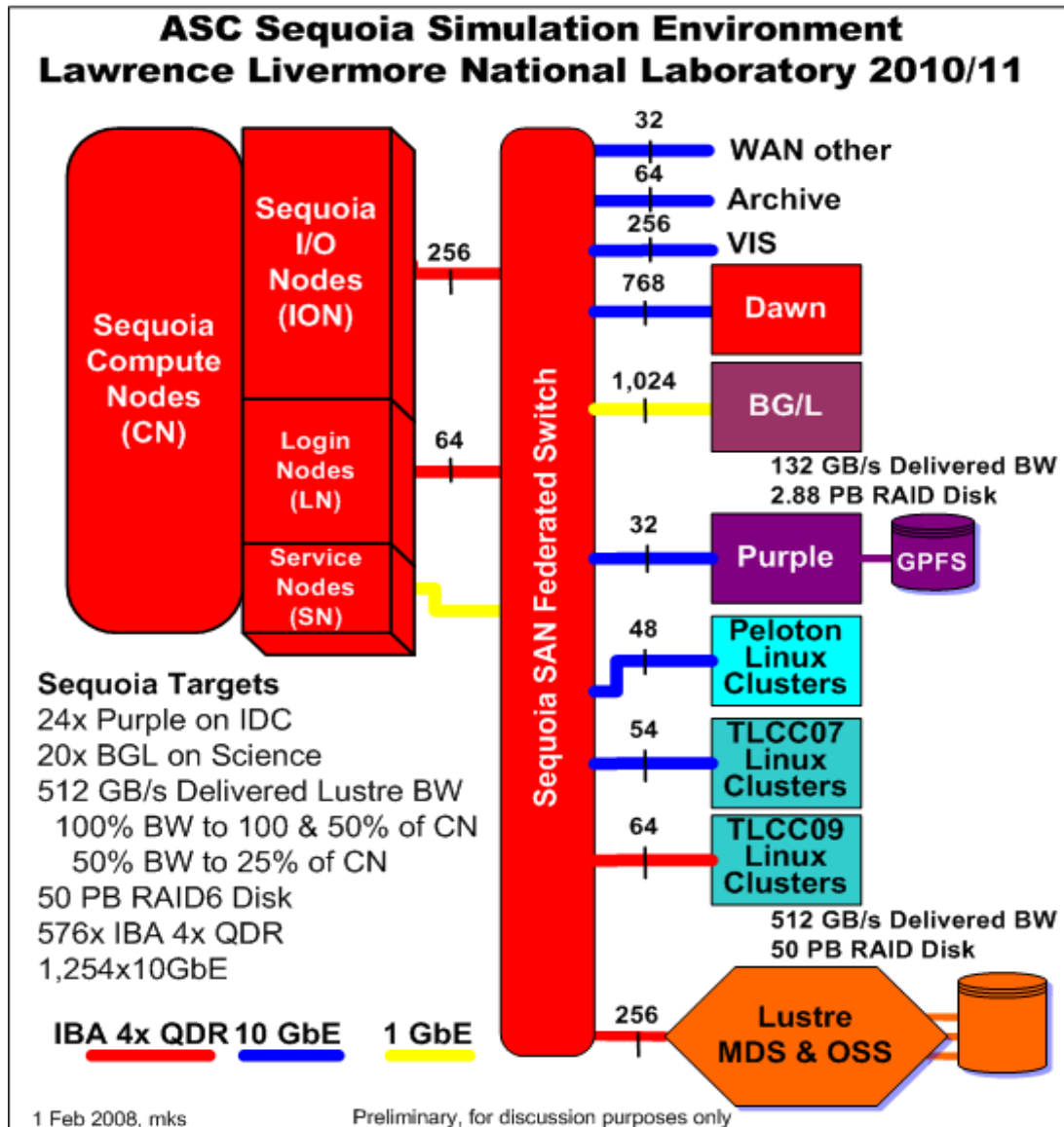


# DOE Office of Science Next System

---

- DOE's requirement for 20-40 PF of compute capability split between the Oak Ridge and Argonne LCF centers
- ORNL's proposed system will be based on accelerator technology includes software development environment
- Plans are to deploy the system in late 2011 with users getting access in 2012

# Sequoia LLNL

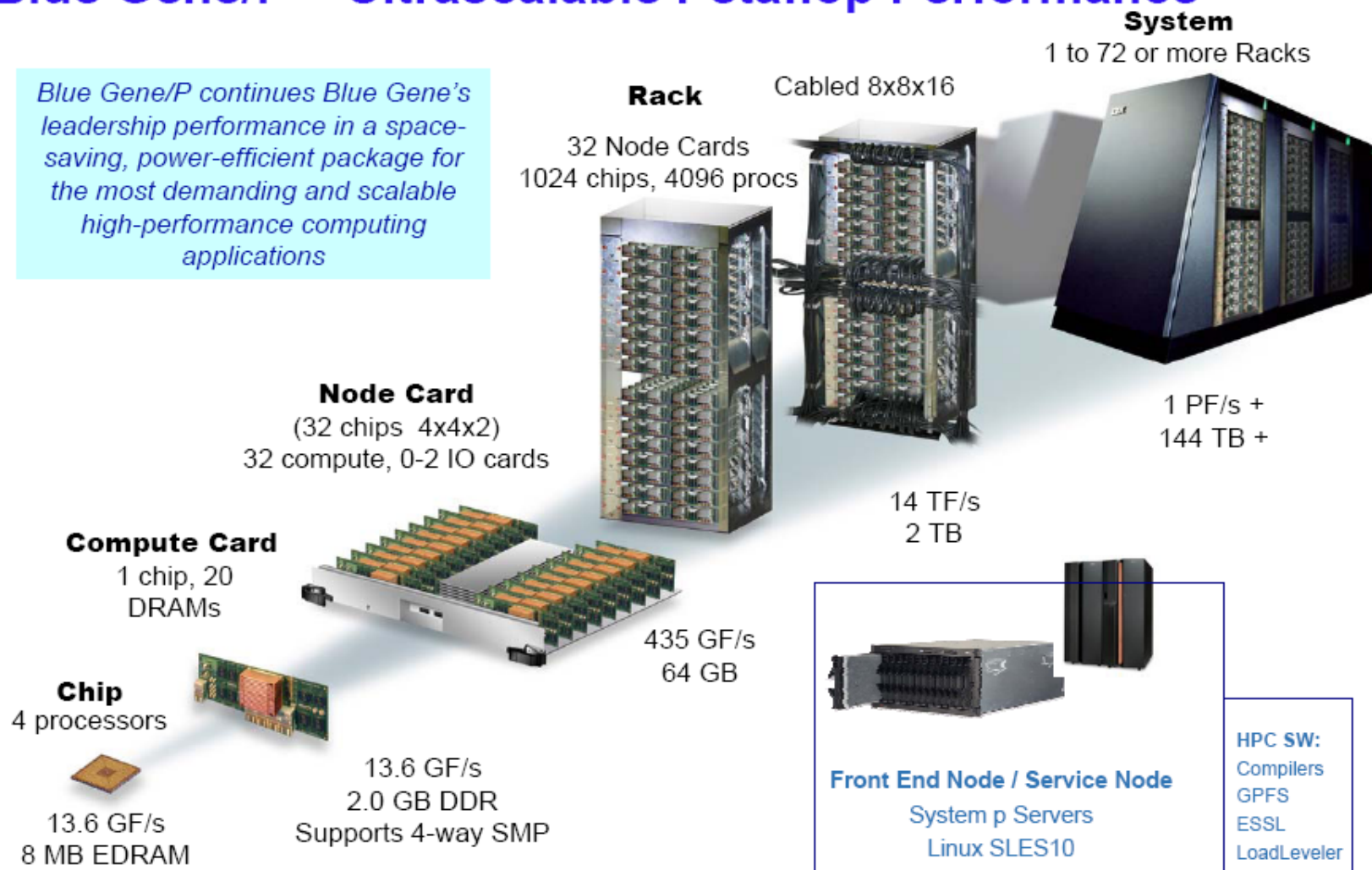


- **Diverse usage models drive platform and simulation environment requirements**
  - Will be 2D ultra-res and 3D high-res Quantification of Uncertainty engine
  - 3D Science capability for known unknowns and unknown unknowns
- **Peak 20 petaFLOP/s**
- **IBM BG/Q**
- **Target production 2011-2016**
- **Sequoia Component Scaling**
  - **Memory B:F = 0.08**
  - **Mem BW B:F = 0.2**
  - **Link BW B:F = 0.1**
  - **Min Bisect B:F = 0.03**
  - **SAN BW GB/:PF/s = 25.6**
  - **F is peak FLOP/s**



## Blue Gene/P – Ultrascalable Petaflop Performance

*Blue Gene/P continues Blue Gene's leadership performance in a space-saving, power-efficient package for the most demanding and scalable high-performance computing applications*





# NSF University of Illinois; Blue Waters

---

Blue Waters will be the powerhouse of the National Science Foundation's strategy to support supercomputers for scientists nationwide

T1	Blue Waters	NCSA/Illinois	1 Pflop <i>sustained</i> per second
T2	Ranger	TACC/U of Texas	504 Tflop/s peak per second
	Kraken	NICS/U of Tennessee	1 Pflops peak per second
T3	Campuses across the U.S.	Several sites	50-100 Tflops peak per second





# NSF University of Illinois; Blue Waters

---

## Blue Waters - Main Characteristics

- **Hardware:**

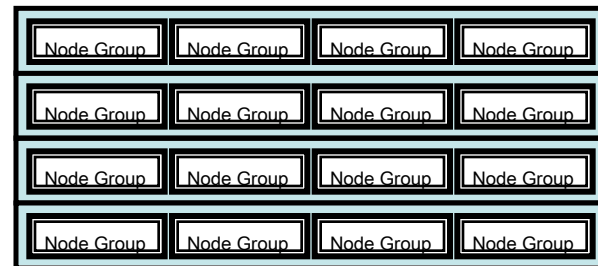
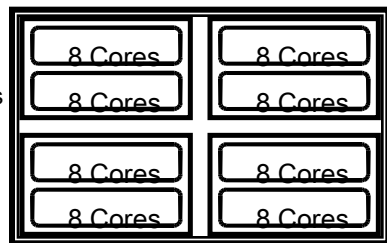
- Processor: IBM Power7 multicore architecture
  - 8 cores per chip, 4 GHz
  - 32 Gflop/s per core; 256 Gflop/s chip
- More than 200,000 cores will be available
- Capable of simultaneous multithreading (SMT)
- Vector multimedia extension capability (VMX)
- Four or more floating-point operations per cycle
- Multiple levels of cache - L1, L2, shared L3
- 32 GB+ memory per SMP, 2 GB+ per core
- 16+ cores per SMP
- 10+ Petabytes of disk storage
- Network interconnect with RDMA technology

# PERCS Hardware Architecture (NCSA System)

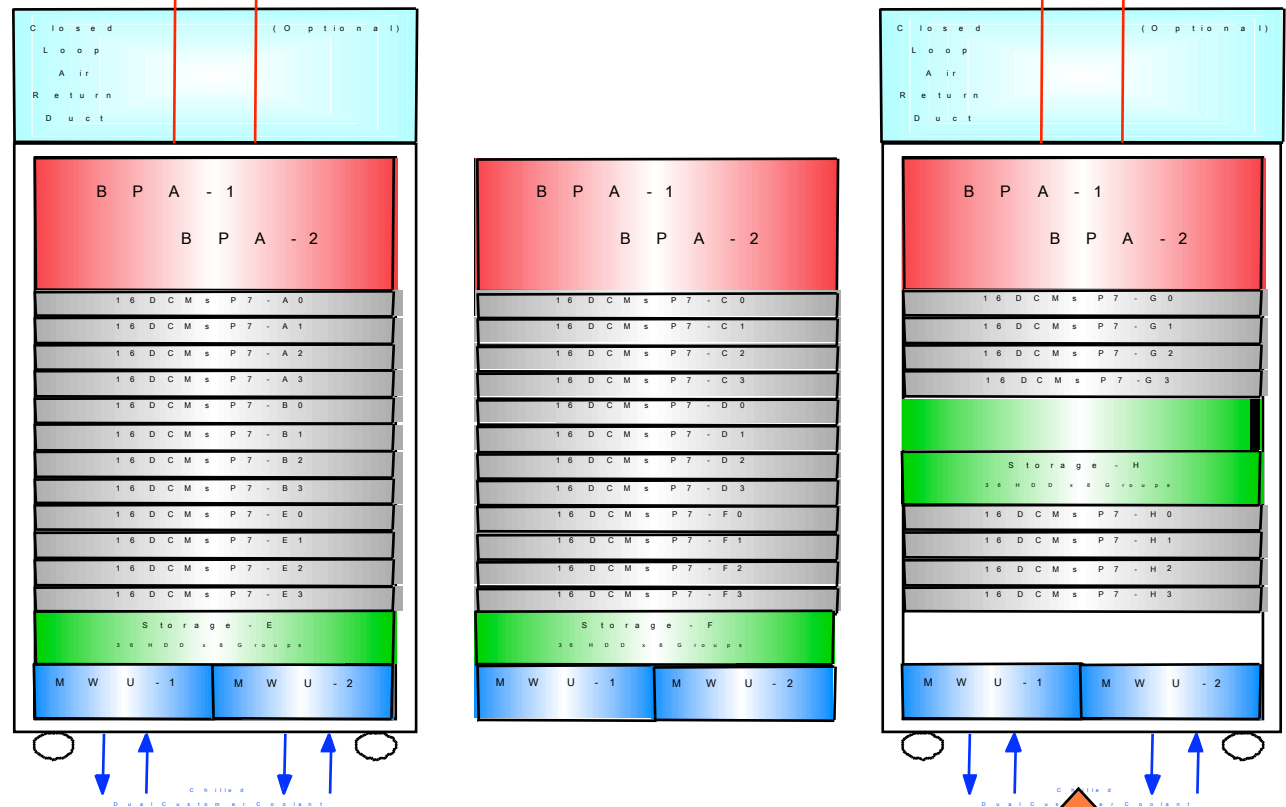
System Building Block  
3 racks, 8 supernodes  
4 disk drawers (288 drives/drawer)  
20 tape drives (situated remotely)  
NCSA system is 38 building blocks

Dual-Chip Module  
2 Power7 chips  
16 cores (8 cores/chip)  
32 GF/core, 256 GF/chip

Node Group  
4 dual-chip modules, 64 cores  
8/16/32/64-way SMP  
128 GB RAM, 2.05TF  
4 node groups per 2U drawer



Supernode  
4x2U drawers, 16 node groups  
1024 cores, 32.3TF  
2 TB mem (2G per core)





# PERCS Hardware (NCSA system)

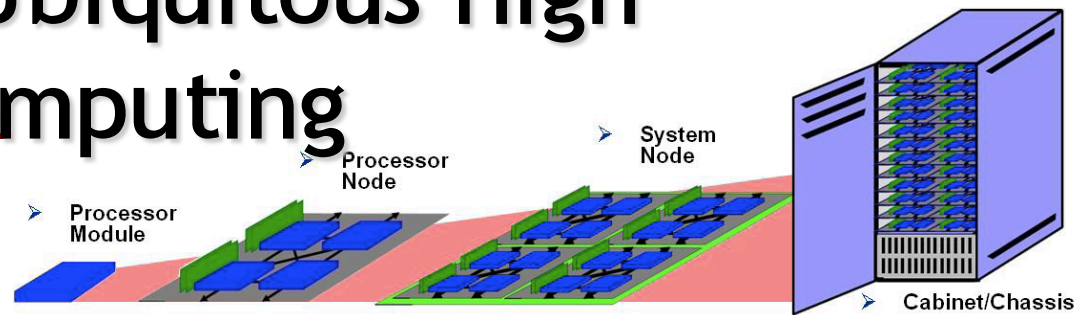
System Peak	<b>10.06 PF</b> <b>38,912 8-way 4.04 GHz POWER7 chips; 45 nm technology</b>
HPCC HPL	<b>8.2 PF (estimate)</b>
Min/Max Number of OS Images	<b>4,864 (64 way) to 38,912 (8 way) Linux or AIX OS images</b>
FLOPs/Core, FLOPs/Chip, FLOPs/Socket, FLOPs/Supernode	<b>32.3 GF per core, 258.6 GF per chip, 517.1 GF per socket, 331 TF/supernode</b>
Threads/Core	<b>4-way SMT</b>
Total Cache Memory	<b>1.3 TB</b>
Total System Main Memory	<b>623 TB, IBM Pulsar buffered DIMMS</b>
Total Main Memory Available to Users	<b>556 TB (38,912 SMPs), 574 TB (4,864 SMPs)</b>
Total Memory Bandwidth	<b>5.0 PB/s (B/F=0.5; L1: B/F=6; L3: B/F=3)</b>
HPCC STREAM	<b>3.10 PB/s (estimate)</b>
Peak Interconnect Bandwidth	<b>1.37 PB/s</b>
Disk Storage	<b>26.3 PB raw, 23.3 usable (not including RAID6+ with spares)</b>
Archival Storage	<b>Up to 1 EB</b>
Total Storage Bandwidth	<b>4.38 TB/s raw, 2.02 TB/s sustained (disk) + 100 GB/s (tape)</b>

Time to Load or Store User Memory from or to Disk	<b>Load: ~5 minutes; store ~10 minutes</b>
Time to Perform Checkpoint/Restart	<b>15-20 minutes (estimate)</b>
Time to Start Full System Job	<b>~5 minutes (estimate)</b>
Total System MTBF	<b>14 days</b>
External Network Bandwidth	<b>440 Gb/s using 44 10 GbE connections</b>
Power	<b>10.3 MW (Average Continuous Power)</b>
Floor Space	<b>114 integrated compute/storage racks occupying 4,452 sq feet</b>
Field Replaceable Unit	<b>Hot swappable drawer with 32 POWER7 chips (256 cores)</b>
Boot Time For Full System	<b>Cold boot: &lt; 2 hours; warm boot: &lt; 1 hour</b>



# DARPA's RFI on Ubiquitous High Performance Computing

- *1 PFLOP/S HPL, air-cooled, single 19-inch cabinet ExtremeScale system.*
- *The power budget for the cabinet is 57 kW, including cooling.*
- *Achieve 50 GFLOPS/W for the High-Performance Linpack (HPL) benchmark.*
- *The system design should provide high performance for scientific and engineering applications.*
- *The system should be a highly programmable system that does not require the application developer to directly manage the complexity of the system to achieve high performance.*
- *The system must explicitly show a high degree of innovation and software and hardware co-design throughout the life of the program.*
- *5 phases;*
  - *1) conceptual designs; 2) execution model; 3) full-scale simulation; 4) delivery; 5) modify and refine.*



# Exascale Computing

- Exascale systems are likely feasible by 2017±2
- 10-100 Million processing elements (cores or mini-cores) with chips perhaps as dense as 1,000 cores per socket, clock rates will grow more slowly
- 3D packaging likely
- Large-scale optics based interconnects
- 10-100 PB of aggregate memory
- Hardware and software based fault management
- Heterogeneous cores
- Performance per watt — stretch goal 100 GF/watt of sustained performance  $\Rightarrow >> 10 - 100$  MW Exascale system
- Power, area and capital costs will be significantly higher than for today's fastest systems

## ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

Peter Kogge, Editor & Study Lead  
Keren Bergman  
Shekhar Borkar  
Dan Campbell  
William Carlson  
William Dally  
Monty Denneau  
Paul Franzone  
William Harrod  
Kerry Hill  
Jon Hiller  
Sherman Karp  
Stephen Keckler  
Dean Klein  
Robert Lucas  
Mark Richards  
Al Scarpelli  
Steven Scott  
Allan Snavely  
Thomas Sterling  
R. Stanley Williams  
Katherine Yelick

September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number FA8650-07-C-7724. This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

### NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation, or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED.



# IESP: The Need

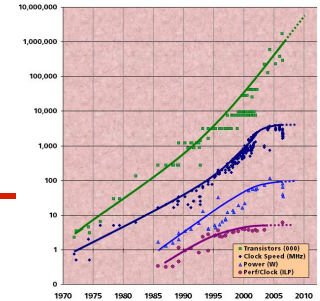
---



- The largest scale systems are becoming more complex, with designs supported by consortium
  - The software community has responded slowly
- Significant architectural changes evolving
  - Software must dramatically change
- Our ad hoc community coordinates poorly, both with other software components and with the vendors
  - Computational science could achieve more with improved development and coordination

# A Call to Action

- Hardware has changed dramatically while software ecosystem has remained stagnant
- Previous approaches have not looked at co-design of multiple levels in the system software stack (OS, runtime, compiler, libraries, application frameworks)
- Need to exploit new hardware trends (e.g., manycore, heterogeneity) that cannot be handled by existing software stack, memory per socket trends
- Emerging software technologies exist, but have not been fully integrated with system software, e.g., UPC, Cilk, CUDA, HPCS
- Community codes unprepared for sea change in architectures
- No global evaluation of key missing components







# IESP Goal

---

Improve the world's simulation and modeling capability by improving the coordination and development of the HPC software environment

Workshops:

**Build an international plan for developing the next generation open source software for scientific high-performance computing**



# Where We Are Today:

---

- ☐ SC08 (Austin TX) meeting to generate interest
  - ☐ DOE's Office of Science funding
  - ☐ US meeting April 6-8, 2009
    - ☐ 65 people
  - ☐ NSF's Office of Cyberinfrastructure funding
  - ☐ European meeting June 28-29, 2009
    - ☐ 70 people
    - ☐ Draft Roadmap
    - ☐ Outline Report
  - ☐ Asian meeting (Tsukuba Japan) October 18-20, 2009
    - ☐ Refine roadmap
    - ☐ Refine Report
  - ☐ SC09 (Portland OR) BOF to inform others
    - ☐ Public Comment
    - ☐ Draft Report presented
- Nov 2008
- Apr 2009
- Jun 2009
- Oct 2009
- Nov 2009



# All of these issues add programming complication

---

- Assertion: data structure design and data motion minimization have
- more impact on performance than instruction ordering
  - But, these are both architecture specific!
- Resilience: DARPA exascale report has component failure 35-39 mins
  - Message delivery failure in MPI-3
  - Dead node detection and recovery
  - Needs to be integrated from the hardware through the application
- Soft error tolerance
  - If we assume any operation can give incorrect results, can we make more robust algorithms?
  - Can we better protect high-order bits?
- Some hardware support libraries are only available in certain programming languages, and some programming models only on certain hardware

# Conclusions

---

- For the last decade or more, the research investment strategy has been overwhelmingly biased in favor of hardware.
- This strategy needs to be rebalanced - barriers to progress are increasingly on the software side.
- Moreover, the return on investment is more favorable to software.
  - Hardware has a half-life measured in years, while software has a half-life measured in decades.
- High Performance Ecosystem out of balance
  - Hardware, OS, Compilers, Software, Algorithms, Applications
    - No Moore's Law for software, algorithms and applications

# Collaborators / Support

Employment opportunities for  
post-docs in the ICL group  
at Tennessee



NVIDIA

Microsoft



The MathWorks



- Top500
  - Hans Meuer, Prometheus
  - Erich Strohmaier, LBNL/NERSC
  - Horst Simon, LBNL/NERSC

Google™

  
  [Advanced Search](#)  
[Preferences](#)  
[Language Tools](#)

[Advertising Programs](#) - [Business Solutions](#) - [About Google](#)

©2007 Google



# If you are wondering what's beyond ExaFlops

---

## Mega, Giga, Tera, Peta, Exa, Zetta ...

$10^3$	kilo
$10^6$	mega
$10^9$	giga
$10^{12}$	tera
$10^{15}$	peta
$10^{18}$	exa
$10^{21}$	zetta

$10^{24}$	yotta
$10^{27}$	xona
$10^{30}$	weka
$10^{33}$	vunda
$10^{36}$	uda
$10^{39}$	treda
$10^{42}$	sorta
$10^{45}$	rinta
$10^{48}$	quexa
$10^{51}$	pepta
$10^{54}$	ocha
$10^{57}$	nena
$10^{60}$	minga
$10^{63}$	luma